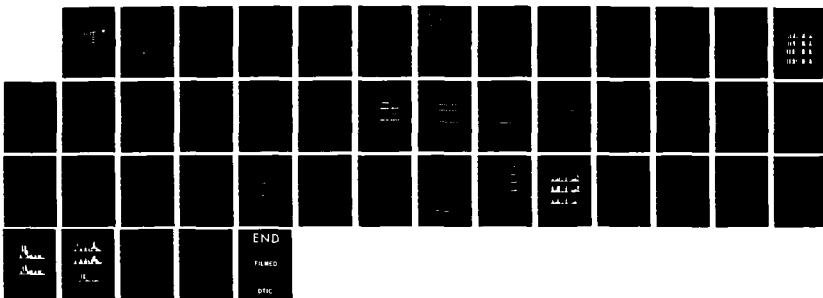
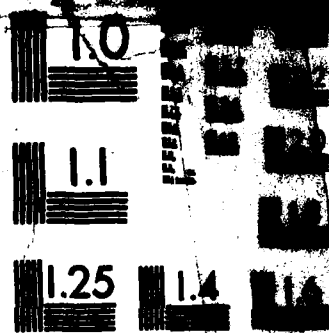


AD-A145 031 IMPROVEMENT OF THE NARROWBAND LINEAR PREDICTIVE CODER 1/1
PART 2 SYNTHESIS IMPROVEMENTS(U) NAVAL RESEARCH LAB
WASHINGTON DC G 5 KANG ET AL. 11 JUN 84 NRL-8799
UNCLASSIFIED SBI-AD-E000 588 F/G 17/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Improvement of the Narrowband Linear Predictive Coder

Part 2—Synthesis Improvements

GEORGE S. KANG AND STEPHANIE S. EVERETT

*Communication Systems Engineering Branch
Information Technology Division*

AD-A145 031

June 11, 1984

DTIC FILE COPY



NAVAL RESEARCH LABORATORY
Washington, D.C.

DTIC
ELECTE

AUG 21 1984

A

Approved for public release; distribution unlimited.

84 08 21 041

AD-A145031

SECURITY CLASSIFICATION OF THIS PAGE

| REPORT DOCUMENTATION PAGE | | | |
|--|--|---|--|
| 1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED | | 1b. RESTRICTIVE MARKINGS | |
| 2a. SECURITY CLASSIFICATION AUTHORITY | | 3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited. | |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | | | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Report 8799 | | 5. MONITORING ORGANIZATION REPORT NUMBER(S) | |
| 6a. NAME OF PERFORMING ORGANIZATION Naval Research Laboratory | 6b. OFFICE SYMBOL (If applicable) Code 7526 | 7a. NAME OF MONITORING ORGANIZATION | |
| 6c. ADDRESS (City, State and ZIP Code) Washington, DC 20375 | | 7b. ADDRESS (City, State and ZIP Code) | |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research | 8b. OFFICE SYMBOL (If applicable) | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER | |
| 8c. ADDRESS (City, State and ZIP Code) Arlington, VA 22217 | | 10. SOURCE OF FUNDING NOS. | |
| | | PROGRAM ELEMENT NO. 61153N | PROJECT NO. RR021-05-42 |
| | | TASK NO. | WORK UNIT NO. DN 280-209 |
| 11. TITLE (Include Security Classification) (See Page ii) | | | |
| 12. PERSONAL AUTHOR(S) Kang, G. S. and Everett, S. S. | | | |
| 13a. TYPE OF REPORT Final | 13b. TIME COVERED FROM TO | 14. DATE OF REPORT (Yr., Mo., Day) 1984 June 11 | 15. PAGE COUNT 45 |
| 16. SUPPLEMENTARY NOTATION | | | |
| 17. COSATI CODES | | 18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) | |
| FIELD | GROUP | SUB. GR. | |
| | | LPC speech synthesis | |
| | | Speech improvements | |
| | | Excitation signal | |
| | | Prediction residual | |
| | | Pitch jitter | |
| | | Output bandwidth expansion | |
| 19. ABSTRACT (Continue on reverse if necessary and identify by block number) | | | |
| <p>The narrowband linear predictive coder (LPC) is widely used in both civilian and military applications. Yet in spite of improvements over the years, it is still not universally accepted by general users. This report examines the weakness of the LPC synthesizer, particularly the excitation signal. Diagnostic Acceptability Measure tests show an increase up to five points. This can be achieved without altering the speech sampling rate, the frame rate, or the parameter coding.</p> | | | |
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED-UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS <input type="checkbox"/> | | 21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED | |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL G. S. Kang | | 22b. TELEPHONE NUMBER (Include Area Code) (202) 767-2157 | 22c. OFFICE SYMBOL Code 7526 |

DD FORM 1473, 83 APR

EDITION OF 1 JAN 73 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE

SECURITY CLASSIFICATION OF THIS PAGE

11. TITLE (*Include Security Classification*) (Continued)

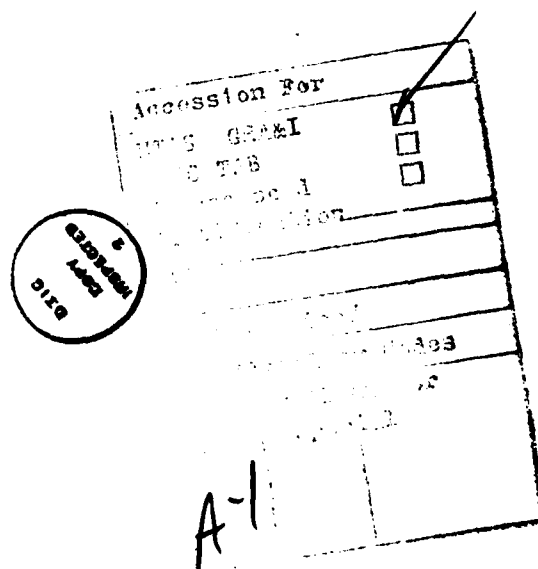
Improvement of the Narrowband Linear Predictive Coder

Part 2—Synthesis Improvements

SECURITY CLASSIFICATION OF THIS PAGE

CONTENTS

| | |
|--|----|
| INTRODUCTION | 1 |
| OVERVIEW OF OUR LPC SYNTHESIS IMPROVEMENTS | 1 |
| BACKGROUND | 3 |
| AMPLITUDE SPECTRUM SHAPING OF THE VOICED EXCITATION SIGNAL | 7 |
| PHASE SPECTRUM SHAPING OF THE VOICED EXCITATION SIGNAL | 12 |
| MODIFIED UNVOICED EXCITATION SIGNAL | 26 |
| EXPANDED OUTPUT BANDWIDTH | 34 |
| CONCLUSIONS | 39 |
| ACKNOWLEDGMENTS | 39 |
| REFERENCES | 39 |



IMPROVEMENT OF THE NARROWBAND LINEAR PREDICTIVE CODER PART 2—SYNTHESIS IMPROVEMENTS

INTRODUCTION

For many years the linear predictive coder (LPC) has been used to convert speech into digital form for secure voice transmission over narrowband channels at low bit rates (less than 5% of the original speech transmission rate). The Navy, as a prime user of narrowband channels for voice communications, has played a significant role in the research and development of LPCs. In 1973 the Navy produced one of the first narrowband LPCs capable of operating in real time. Since 1978 the Navy has been the Department of Defense's (DoD's) technical agent for the development of LPCs intended for triservice tactical use.

Previously [1], we presented our efforts on LPC analysis improvements. The objective of that investigation was to improve the narrowband LPC performance by modifying the LPC analysis without increasing the data rate (2400 bits per second (b/s)) and without violating the interoperability requirements—such as the speech sampling rate and the parameter encoding format—currently adopted by DoD. We chose to work within the confines of these interoperability requirements because they will soon be established as the military standard (MIL-STD-188-113) or the federal standard (FED-STD-1015), and it was hoped that our efforts could benefit the narrowband LPC currently under development for DoD use.

In this report we present our efforts on LPC *synthesis* improvements as the second part of this two-part series. The objective of this investigation is to improve the narrowband LPC performance by modifying the LPC synthesis by using only the data transmitted by the standard DoD narrowband LPC.

OVERVIEW OF OUR LPC SYNTHESIS IMPROVEMENTS

Figure 1 shows that the narrowband LPC synthesizer has three functional blocks: (a) the synthesis filter, (b) the excitation signal generator, and (c) the postsynthesis processor. As we discuss later, the excitation signal generator and the postsynthesis processing are the weakest links in the narrowband LPC synthesizer; we therefore concentrate on these two areas in this report. Three of the four improvements presented involve the excitation signal; the remaining one involves the postsynthesis processing. We do not present any items related to improvement of the synthesis filter because it is basically constrained by the DoD interoperability requirements. The following is an overview of the four improvements discussed in this report.

Amplitude Spectrum Shaping of the Voiced Excitation Signal

The conventional excitation signal used to generate voiced speech is simply an impulse waveform (or any other fixed waveform with a flat amplitude spectrum) which is repeated at the pitch rate. The use of such an excitation would be logical if the LPC analysis filter completely removed speech resonant frequency components so that the prediction residual had a flat amplitude spectral envelope. In actuality, the prediction residual retains a considerable amount of speech resonant frequency components because of limitations inherent in the linear predictive analysis (i.e., the all-pole modeling of the speech and the use of a limited number of filter weights). Therefore, to generate more natural-sounding speech, the narrowband LPC excitation signal should contain resonant frequencies similar to

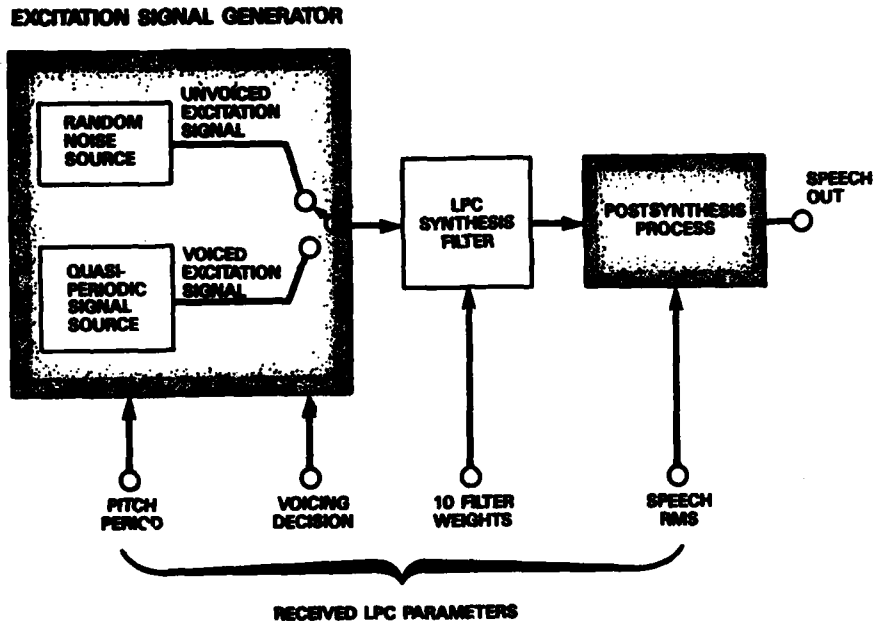


Fig. 1 — Block diagram of the narrowband LPC synthesizer. The shaded blocks are those items we have modified as discussed in this report.

those in the prediction residual. We present a way of introducing these resonant frequencies into the conventional narrowband excitation signal for voiced speech. The amplitude spectrum shaping of the voiced excitation signal produced a 5.2-point improvement in the speech quality as evaluated by the Diagnostic Acceptability Measure (DAM) [2]. This indicates that the resulting speech quality is comparable to that of a voice processor operating at 9600 b/s, or four times the data rate of the narrowband LPC.

Phase Spectrum Shaping of the Voiced Excitation Signal

The individual waveform of the conventional voiced excitation signal repeats exactly from one pitch cycle to the next. In contrast, the prediction residual rarely repeats exactly from one pitch cycle to the next. This is due to irregularities in vocal cord movement and turbulent air flow from the lungs during the glottis-open period of each pitch cycle. The extreme regularity of the LPC excitation signal causes the synthesized speech to sound machinelike and tense. To reduce this effect, pitch epoch variations and period-to-period waveform variations may be conveniently realized by introducing phase jitter in the waveform. We present a new expression for the voiced excitation signal and specify the phase jitter characteristics. Use of this phase spectrum shaping in the voiced excitation signal increased overall quality DAM scores by 4.7 points for male speakers and 5.0 points for female speakers.

Modified Unvoiced Excitation Signal

The conventional excitation signal for generating unvoiced speech is simply random noise with a uniform or Gaussian amplitude distribution. Such an excitation produces satisfactory nonabrupt unvoiced sounds, or continuants, such as /f/, /s/, /sh/, and /th/. As expected, the prediction residuals for these sounds are random, with an approximately Gaussian amplitude distribution. On the other hand, the prediction residuals for abrupt consonants such as /k/, /t/, and /ch/ are spiky and irregular, especially in the burst or onset portion of the sound. Therefore the satisfactory production of these sounds requires an excitation signal consisting of random noise with at least one large spike at the onset. Without this large spike, a synthesized stop consonant usually sounds more like a continuant.

We present a new form of the unvoiced excitation signal. Although similar to the conventional unvoiced excitation for the generation of nonabrupt unvoiced sounds, our excitation signal generates randomly spaced spikes if the speech root-mean-square (RMS) value changes sharply from one unvoiced frame to another. This modified unvoiced excitation signal enhances the reproduction of unvoiced plosives without affecting the reproduction of nonabrupt unvoiced sounds.

The use of the modified unvoiced excitation signal improved the overall Diagnostic Rhyme Test (DRT) [3] score of the LPC by 3.6 points for three female speakers. Significantly, the partial score for discriminating abrupt vs nonabrupt unvoiced sounds was improved by 14.4 points, implying that we have properly identified a major weakness in the unvoiced excitation signal and generated a solution to correct it.

Expanded Output Bandwidth

Contrary to convention, the output bandwidth of a voice processor need not be the same as the input bandwidth. According to our experimentation, synthesized speech is much brighter and often more intelligible when the output bandwidth is made greater than the input bandwidth. To accomplish this in the narrowband LPC without altering the data rate, we folded the frequency contents of synthesized speech between 2 and 4 kHz upward at 4 kHz to make an output bandwidth of 6 kHz, rather than the usual 4 kHz. This results in more natural fricative sounds and sharper stop consonants. Although this also generates weak extraneous formants in the upperband regions of voiced speech sounds, it does not affect their intelligibility, and in fact adds brightness to their tonal quality. Test results show that the extended output bandwidth produces a 2.5-point increase in overall quality as measured by the DAM.

BACKGROUND

Over the years numerous voice processors have been developed for operational use, including pulse code modulators (PCM) at 18.75 and 50 kilobits per second (kb/s), continuously variable slope delta (CVSD) modulators at 16 and 32 kb/s, adaptive predictive coders (APC) at 6.4 and 9.6 kb/s, and the narrowband LPC and a channel vocoder at 2.4 kb/s. Today the most commonly used data rates are 2.4, 9.6, and 16 kb/s.

The narrowband LPC operating at 2.4 kb/s is becoming a vital part of the DoD voice communication system because it can provide adequate communicability in less than favorable operational environments. For example, it can transmit speech over narrowband channels with a bandwidth of approximately 3 kHz, such as high frequency (HF) channels, unequalized telephone lines, or fieldwires. Transmission over HF channels, which the Navy often relies on, requires a simple low-power transmitter operable in shipboard, airborne, shelter, and vehicular platforms.

The narrowband LPC can also transmit speech more reliably over the Navy FLEETSATCOM channels than can higher data rate voice processors. Because the fixed power at the satellite relay makes the signal-to-noise ratio at the receiver inversely proportional to the data rate, the low data rate of the 2.4 kb/s LPC provides a less noisy speech signal.

Furthermore, the narrowband LPC provides better survivability in the presence of man-made or natural disturbances in the transmission channel since there are more narrowband channels available for rerouting (such as public and DoD telephone lines). In addition, the 2.4 kb/s narrowband LPC actually yields higher intelligibility scores than some higher rate voice processors in certain high-noise environments. For example, in a shipboard platform the average DRT score for the narrowband LPC is 87.2, whereas it is only 80.0 for the 9.6 kb/s APC.

Because of these advantages, the use of the narrowband LPC is expected to become more widespread in the future. Although the narrowband LPC may outperform higher rate voice processors in less favorable operational conditions, it is still inferior when operated in a quiet environment. In general, the intelligibility of narrowband LPC speech is moderately good. The average overall DRT scores are about 89 for male talkers and about 86 for female talkers, which compare favorably with those of the 9.6 kb/s APC (91 for both male and female talkers). However, the speech quality of the LPC is notoriously poor. For example, the Composite Acceptability Estimate (CAE) of the Diagnostic Acceptability Measure (DAM) for the narrowband LPC is about 6 points lower than that of the APC for male talkers, and 9 points lower for female talkers.

Weaknesses of the Narrowband LPC Synthesizer

The synthesis procedure in the narrowband LPC is partly to blame for the deficiency in speech quality mentioned above because the model used to generate the speech is simple and unrealistic. The narrowband LPC excitation signal is based on the assumption that all speech can be generated by using either a purely periodic (voiced) excitation, or a purely random (unvoiced) excitation. The weakness of this model becomes evident when it is compared with the prediction residual representing the ideal excitation signal for the LPC analysis/synthesis system. The prediction residual, unlike the narrowband LPC excitation signal, is not always periodic, even when the input speech is a sustained vowel. Likewise, the prediction residual is not always random when the input speech is unvoiced. Most importantly, the prediction residual is a sample-by-sample quantity that cannot be closely approximated by a signal which is regenerated by using a limited number of frame-by-frame parameters as is the case with the narrowband LPC excitation signal.

One way of improving the excitation signal would be to transmit the prediction residual itself, as in the APC or the Navy Multirate Processor (MRP) [4]. However, to do this requires a data rate of at least 9.6 kb/s. Another way to improve the excitation signal would be to create a multipulse signal to minimize the perceptual difference between the unprocessed and the synthetic speech [5]. Still, the required data rate is well in excess of 2.4 kb/s.

Because any improvements to the narrowband LPC must be interoperable with the standard DoD narrowband LPC, we do not propose to use a radically different excitation signal. We do, however, propose to use a more general form of the excitation signal source from which either the voiced or the unvoiced excitation signal or a hybrid signal resembling both, may be generated. This modified excitation signal source has more control variables than the conventional source, allowing more freedom in specifying its characteristics.

Modified Excitation Signal Source

The conventional excitation signal is divided into two mutually exclusive parts: a broadband repetitive signal to generate voiced speech and a broadband random signal to generate unvoiced speech. The choice between the two excitation signals is determined by the (binary) voicing decision; the repetitive rate of the voiced excitation signal is governed by the pitch frequency.

In contrast, our modified excitation signal is not rigidly divided into two classes—the voiced excitation signal contains some random components, and, likewise, the unvoiced excitation signal contains some deterministic components. This hybrid form of excitation signal is much closer to the actual voicing excitation than is the conventional divided signal. As we show, the presence of these complementary components improves the naturalness and quality of the synthesized speech.

In essence, the conventional excitation signal is a stationary model of our excitation signal. The conventional signal is generated under the assumptions that (a) the amplitude spectrum is flat and

time-invariant, (b) the phase spectrum of the voiced excitation signal is a time-invariant function of frequency, and (c) the phase spectrum of the unvoiced excitation signal has a probability function that is time invariant. These assumptions make it possible to generate a replica of the voiced excitation signal which can be stored in memory and read out sequentially at every voiced pitch epoch. Similarly, unvoiced excitation signal samples are read out randomly from a table containing uniformly distributed random numbers.

In our modified excitation signal we do not use "canned" samples with invariant characteristics. Instead we generate new excitation signal samples at each pitch epoch, or at a fixed time interval if the speech is unvoiced, based on the updated amplitude and phase spectra of the excitation. This excitation signal is based on the Fourier series; thus the m th excitation sample $e(i)$ is given by

$$e(i) = \sum_{k=0}^K a(k) \cos \left[\left(\frac{2\pi k}{I} \right) i + \phi(k) \right], \quad 1 \leq i \leq I \quad (1)$$

where $a(k)$ and $\phi(k)$ are the k th amplitude and phase spectral components, respectively, I is the number of excitation signal samples, and K is the number of amplitude or phase spectral components. The quantity K is related to I by

$$K = \begin{cases} \frac{I}{2} + 1 & \text{if } I \text{ is even} \\ \frac{I+1}{2} & \text{if } I \text{ is odd.} \end{cases} \quad (2)$$

Equation (1) is the most general form of the excitation signal. It represents the excitation signal not only for the narrowband LPC, but also for the wideband LPC as in the previously mentioned Navy MRP [4]. In the MRP, the quantity I in Eq. (1) is the frame width, and both the amplitude and phase spectral components, $a(k)$ and $\phi(k)$, are derived from the actual prediction residual. Thus, the resulting speech quality (at 16 kb/s) is excellent.

The conventional narrowband LPC excitation signal may also be expressed by Eq. (1). In this representation, the voicing decision is mapped onto the phase spectrum. Thus, the conventional excitation signal in the form of Eq. (1) has two different phase spectra since it is controlled by a two-state voicing decision. Table 1 gives the general characteristics of these two types of phase spectra. As we will show, these correspond to the stationary parts of the phase spectrum of our modified excitation signal for the respective voicing modes. The amplitude spectrum is, of course, flat and time invariant.

Our modified excitation signal will have spectral properties as described in Table 1. The methods for generating these characteristics and the rationale behind them are discussed in a subsequent section of this report.

The duration of the narrowband LPC excitation signal is denoted by I in Eq. (1). If the speech is voiced, the quantity I corresponds to the length of the pitch period as received by the synthesizer. If the speech is unvoiced, there is by definition no pitch period, so we assign a fixed time interval, similar to a pitch period, to periodically renew the unvoiced excitation signal and to periodically interpolate the LPC parameters.

The unvoiced excitation signal is dispersed over the entire time interval because its phase spectral components are randomly distributed (see Table 1). However, this is not the case with the voiced excitation signal. For example, if we assume that the amplitude spectrum is flat and the phase spectrum is a linear function of frequency, then the resulting voiced excitation signal is an impulse, meaning that

Table 1—Summary of Narrowband LPC Excitation Signal Parameters

| Parameters | | Conventional Narrowband LPC Excitation Signal | Our Modified Narrowband LPC Excitation Signal |
|---------------------------|-----------------|--|---|
| Amplitude Spectrum $a(k)$ | | Frequency-independent and time-invariant (assigned parameter) | With weak resonant frequencies updated pitch-synchronously |
| Phase Spectrum $\phi(k)$ | Voiced Speech | A nonlinear function of frequency, and time-invariant (assigned parameter) | A quadratic function of frequency, with frequency-dependent phase jitters |
| | Unvoiced Speech | N/A ^a | A stationary random process with a uniform distribution between $-\pi$ and π radians, superimposed by amplitude-weighted, randomly spaced pulses. |
| Signal Duration (I) | | Pitch period (received parameter) | Pitch period (received parameter) |

^aMost commonly, the conventional unvoiced excitation signal is read out randomly from a table containing uniformly distributed random numbers. Its phase spectrum cannot be expressed conveniently in terms of Eq. (1).

only one out of I excitation samples is nonzero. The spread of the voiced excitation signal is dependent on the phase spectrum. We present a preferred phase spectrum for the voiced excitation signal in a later section of this report.

Test and Evaluation of Synthesized Speech

Even though there is no "speech quality meter" that automatically indicates the quality of synthetic speech, tests using known quality evaluation methods, such as the DAM test, are time-consuming, particularly when the processor does not run in real time. For this reason, researchers often perform so-called "informal listening tests." This method can indicate speech quality when done by using naive listeners, but such tests can be rather misleading when the researchers themselves act as listeners because their ears have been conditioned to the electronic accents of their own voice processors. Furthermore, the aspect of speech they are trying to improve may be easily heard by the researchers but imperceptible to casual or untrained listeners. Therefore, it is essential to use established test methods for quality evaluation.

However, quality evaluation using established methods is not all that is needed; one must check carefully to be sure that a change in one aspect of the voice processing does not degrade another area. For example, filtering out the synthesized speech components below approximately 250 Hz produces a more spectrally balanced sound for the narrowband LPC. Many listeners prefer this because the absence of a heavy bass component makes the upper frequency contents more noticeable and intelligible. However, such an alteration must be tested for potentially adverse effects on pitch and voicing estimation when the LPC is operated in tandem with another narrowband LPC. Likewise any modification to one aspect of the speech must be tested for effects on other aspects. Frequently an improvement in subjective speech quality degrades the measured speech intelligibility.

In this report we have chosen to use evaluation methods that are sensitive to the specific aspects of speech we are trying to improve. For example, the Diagnostic Rhyme Test (DRT), which measures the intelligibility of initial consonants, would not be the best method to use for evaluating the quality of

synthesized speech. A much better evaluation could be made by using a method such as the Diagnostic Acceptability Measure (DAM) that is specially designed to be sensitive to speech quality.

With the DAM, a system is rated by using 12 phonetically balanced 6-syllable sentences from each talker. A listener hears the 12 sentences as a group, and then rates the overall voice quality on 21 separate rating scales which describe the speech quality, the background noise, and the total effect of the voice signal (e.g., nasal, unnatural, crackling, intelligible). All the scales are combined into an overall composite score. Also, a number of diagnostic scales related to the perceptual quality of the speech signal and the background noise (such as fluttering, muffled, hissy) are computed based on various subsets of the test scales.

Both the DAM and the DRT use standard tape recordings and are scored by Dynastat, Inc. in Austin, Texas, which maintains a stable crew of trained listeners. In this way we may compare our results with those obtained at different times by other researchers. Because these tests measure different aspects of the speech, both have become indispensable tools for evaluating the quality and intelligibility of voice processing systems in the DoD community.

Past Improvements to the LPC Synthesis

It has been nearly a decade since the Navy and others first implemented the narrowband LPC for real-time operation. Since then there have been many improvements related to the narrowband LPC synthesis. The current DoD standard narrowband LPC has incorporated many of the earlier changes developed both by DoD scientists and by R&D firms for their DoD sponsors [6,7]. All these improvements are supported by rational principles as outlined in their respective articles and reports. The features do not adversely affect other aspects of the narrowband speech and we recommend them for any narrowband voice processor. They include the following:

- the use of pitch-synchronous parameter interpolation to make the synthetic speech sound cleaner,
- fixed-power excitation and postsynthesis amplitude calibration to enhance computational accuracy,
- the use of a time-dispersed voiced excitation signal to reduce the speech dynamic range and improve the tandem performance with a continuously variable slope delta (CVSD) processor,
- the use of the speech power, rather than the excitation signal power, as an amplitude parameter to eliminate speech amplitude variations caused by transmission errors in LPC coefficients, and
- nonlinear interpolations of LPC coefficients and the amplitude parameter to highlight sudden speech transitions and make them sound crisper.

Despite all these improvements, the speech quality of the narrowband LPC is still somewhat poor, and the intelligibility of female voices remains lower than that of male voices. This report addresses improvements in these areas.

AMPLITUDE SPECTRUM SHAPING OF THE VOICED EXCITATION SIGNAL

The amplitude spectrum of the synthesized speech is the product of the amplitude spectrum of the excitation signal and the frequency response of the synthesis filter. Thus the quality of the synthesized speech is directly dependent on both these factors. Our objective in this section is to determine the best amplitude spectrum of the excitation signal to use in the narrowband LPC in an effort to

generate the highest quality synthetic speech without compromising the DoD interoperability requirements.

In the conventional narrowband LPC the amplitude spectrum of the excitation signal is always flat, both for the voiced and the unvoiced excitations (i.e., $a(k)$ is a nonnegative constant for all k s in Eq. (1)). However, in looking at the prediction residual as the ideal excitation signal for the LPC, we notice that its amplitude spectrum is not flat at all, especially for voiced speech.

The prediction residual for voiced speech contains a considerable number of resonant frequency components, similar to those in the original speech but lower in intensity (Figs. 2(a) and 2(b)). The presence of these resonant frequencies makes the prediction residual itself highly intelligible. In fact, an average DRT score of 83.5 was obtained by using only the prediction residual for a set of three male speakers (one speaker scored as high as 87.0). Without similar resonant frequency components in the excitation signal, the synthesized speech tends to sound fuzzy and somewhat lacking in clarity.

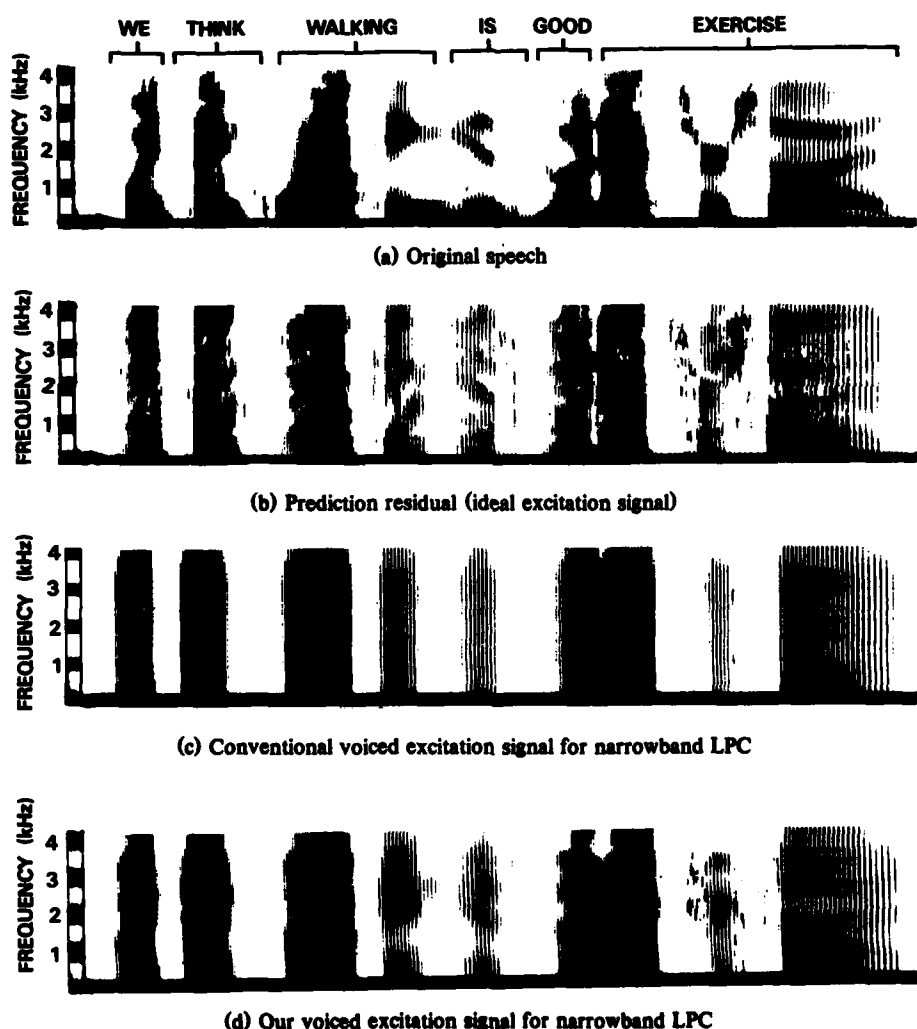


Fig. 2 — Spectra of original speech and LPC excitation signals. The prediction residual contains a considerable number of resonant frequency components unfiltered by the LPC analysis filter; the conventional voiced excitation signal contains no resonant frequencies. Our voiced excitation signal has weak traces of resonant frequencies similar to those of the prediction residual, making the synthesized speech sound more natural.

Resonant Frequencies in the Prediction Residual

In the narrowband LPC the task of the linear predictive analysis is to represent the talker's vocal tract in the form of an all-pole filter. The transfer function of the LPC analyzer transforms the speech waveform to the prediction residual waveform. Thus the residual spectrum $R(z)$, stated in terms of the speech spectrum $E(z)$, is

$$R(z) = \left[1 - \sum_{n=1}^N \alpha_n z^{-n} \right] E(z). \quad (3)$$

The spectral envelope of the residual is flat (i.e., $R(z)$ is a constant) only when the speech spectral envelope is represented perfectly by the all-pole spectrum $H(z)$ expressed by

$$H(z) = \frac{1}{1 - \sum_{n=1}^N \alpha_n z^{-n}} \quad (4)$$

$$= \frac{1}{\prod_{n=1}^{N/2} (1 - z_n z_n^{-1}) (1 - z_n^* z_n^{-1})} \quad (5)$$

where $H(z)$ is equal to the transfer function of the LPC synthesizer, α_n is the n th prediction coefficient, and (z_n, z_n^*) is a complex conjugate pair.

Because of the complex nature of the speech spectrum, the residual spectral envelope $R(z)$ is rarely flat. This is caused in part by the presence of antiresonant components (zeros) in the speech waveform which will not be greatly affected by the LPC analysis filter. Figure 2 illustrates that the prediction residual also contains considerable resonant frequency components not removed by the analysis filter. There are two major reasons for this. First, the magnitudes of the resonant peaks of an all-pole filter, such as the LPC synthesis filter, are dependent on the pole locations (see Eq. (5)); they cannot be independently controlled as they can in a parallel formant synthesizer. In other words, for a given set of pole locations, the magnitudes of the resonant peaks are predetermined and cannot be altered without actually shifting the poles. We have observed that the formant amplitudes in the LPC synthesizer are often lower than those of the actual speech. The greater the magnitude of the original formants, the stronger the resonant frequency components in the prediction residual. Therefore a voice with unusually intense formant frequencies will not be reproduced well by the narrowband LPC unless the excitation signal is augmented with formant frequencies similar to those in the prediction residual.

The second reason why the prediction residual contains considerable resonant frequencies is due to the quantization of the filter coefficients which tends to reduce the spectral peaks attained by an all-pole filter (Fig. 3). This reduction is partly due to the clipping of LPC coefficients by the LPC quantizer. Again, the differentials in the spectral peaks will appear as formant frequencies in the prediction residual. (Figure 3 is based on the coefficient quantization rule for the DoD standard narrowband LPC, but all other parameter quantization rules designed for the 2.4 kb/s LPC produce similar results.)

When the resonant frequency components in the prediction residual are not present in the excitation signal, the synthesized speech lacks clarity. Because the amplitude spectrum of the conventional voiced excitation signal is flat (Fig. 2(c)) the synthesized formants are noticeably muddier than those in the original speech. We have therefore developed a voiced excitation signal containing resonant frequencies which improves the quality of the synthesized speech. Figure 2(d) shows that these resonant frequencies are similar to those contained in the prediction residual.

Earlier Experimentation with Amplitude Shaping

We observed resonant frequencies in the prediction residual as early as 1972 when we first implemented a narrowband LPC based on the flow-form LPC implementation [8]. Unlike the block-form

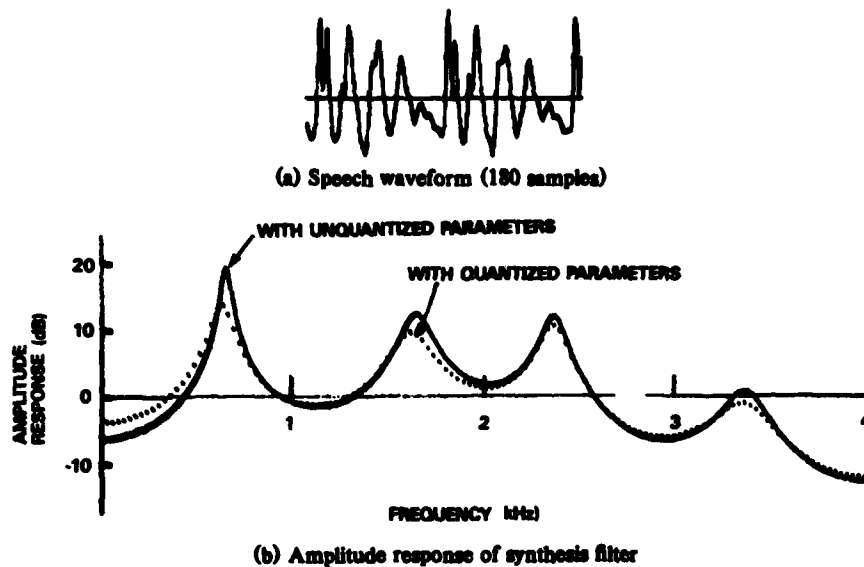


Fig. 3 — Effect of LPC coefficient quantization on the amplitude response of the synthesis filter. Quantization of LPC coefficients results in a reduction of resonant peaks in the synthesis filter.

LPC implementation [6,7], which is often employed because it requires fewer computational steps, the flow-form LPC analysis generates the prediction residual as a by-product of the filter coefficient estimation. We were surprised to find that the prediction residual contained significant resonant frequencies (see Fig. 7 of Ref. 8), and was highly intelligible. We realized that narrowband LPC speech could best be improved by introducing some of these resonant frequencies into the excitation signal.

We investigated methods of shaping the amplitude spectrum of the conventional LPC excitation signal in 1975. An experimental 3.6 kb/s LPC system computed eight additional LPC coefficients from the prediction residual and encoded them into 1.2 kb/s. These eight coefficients were then transmitted along with the conventional 2.4 kb/s LPC data. The sound quality of this 3.6 kb/s LPC was noticeably better than that of the conventional 2.4 kb/s LPC—it was clearer, less muffled, and allowed better speaker recognition. Since we are limited to 2.4 kb/s in the current investigation, we developed a way to achieve similar improvements in speech quality *without* transmitting any additional data derived from the prediction residual. This is a theoretical impossibility; however, an approximate shaping of the excitation signal is possible because the resonant frequencies in the prediction residual track closely with those of the original speech (see again Fig. 2).

Amplitude Spectrum Modification of the Voiced Excitation Signal

Since we are concerned here only with the resonant frequencies in the excitation signal, and not with the antiresonances, the most convenient form of spectral representation is the all-pole spectrum. Thus, let the amplitude spectrum of the modified excitation signal be expressed by

$$A(z) = \frac{1}{1 - \sum_{n=1}^N \gamma_n z^{-n}} \quad (6)$$

where γ_n is the n th prediction coefficient. Ideally γ_n would be obtained from the prediction residual. As noted from Eq. (6), the amplitude spectrum of the modified excitation signal is similar in form to the LPC synthesis filter $H(z)$:

$$H(z) = \frac{1}{1 - \sum_{n=1}^N \alpha_n z^{-n}} \quad (4)$$

where α_n is the n th prediction coefficient obtained from the speech.

While α_n is available at the narrowband LPC receiver, γ_n , which is needed for the amplitude spectral modification, is not. We must therefore approximate γ_n from α_n as best we can. To do this, we exploit two observations.

The first is that the predominant resonant frequencies of the prediction residual track closely with those of the original speech, as illustrated in Fig. 2. This is why the prediction residual is so intelligible. While the prediction residual has extraneous resonant frequencies not found in the original, omission of these does not seem to have a significant impact on the output speech. However the resonant peaks in the prediction residual are nearly equalized, unlike those of the original speech. Thus the all-pole spectrum of the prediction residual may be approximated by the all-pole spectrum of the speech with a reduced feedback gain:

$$\hat{A}(z) = \frac{1}{1 - G \sum_{n=1}^N \alpha_n z^{-n}} \quad G < 1 \quad (7)$$

where α_n is the n th prediction coefficient of the speech available at the LPC synthesizer. The factor G is related to the overall reduction pole moduli. Since the root loci of $\hat{A}(z)$ do not lie along the radial direction there will be a slight but insignificant shift in the frequency of the resonant peaks.

The second observation is that the residual formant peaks become smaller as the prediction residual becomes more random. This occurs with front vowels, murmurs and nasals, where the speech waveform may be well approximated by one or two exponentially decaying sinusoidal functions. For these speech waveforms the efficiency of the linear prediction is fairly high, so that the residual RMS is relatively small for a given speech RMS. Thus, it is natural to assume that the modulus reduction factor is proportional to the ratio of the residual RMS to the speech RMS, namely

$$G = G' \sqrt{\prod_{n=1}^N (1 - w_n^2)} \quad (8)$$

where G' is the proportionality constant yet to be determined, the factor under the radical is the ratio of the residual RMS to the speech RMS, and w_n is the n th reflection coefficient received by the narrowband LPC. (Note that the current narrowband LPC transmits reflection coefficients as the synthesis filter weights. The prediction coefficients are obtained through transformation of the reflection coefficients at the receiver.)

The proportionality constant G' in Eq. (8) is estimated by minimizing the mean-square difference between $A(z)$ of Eq. (6) and $\hat{A}(z)$ of Eq. (7). We chose the frequency-domain computational approach because it enabled us to exclude the effect of frequency components below 150 Hz which were not audible at the narrowband LPC output. We used approximately 1200 frames of male and female voiced speech samples to obtain a preferred value for G' . Not surprisingly, Table 2 shows that G' varies from speaker to speaker. According to this table, a reasonable choice for G' would be somewhere around 0.25, even though, from listening to processed speech while varying G' from 0 to 1.0, it appears that there is a broad range of acceptable values for G' .

The excitation spectrum defined by Eq. (7) may be incorporated in the narrowband LPC in two ways: one is a direct method in which the amplitude spectral components in the excitation signal model in Eq. (1) are made equal to the amplitude spectrum of Eq. (7); the other is an indirect method in which the amplitude spectral components in Eq. (1) are constants, but the amplitude spectrum is

Table 2—Statistics of Proportionality
Constant Used in Eq. (8)

| Speakers | Mean Value | Standard Deviation |
|----------|------------|--------------------|
| Female | 0.394 | 0.157 |
| | 0.366 | 0.165 |
| | 0.299 | 0.141 |
| | 0.182 | 0.052 |
| | 0.135 | 0.027 |
| | 0.122 | 0.004 |
| Male | 0.366 | 0.175 |
| | 0.325 | 0.181 |
| | 0.295 | 0.117 |
| | 0.204 | 0.101 |
| | 0.185 | 0.065 |
| | 0.167 | 0.062 |

Note: For each speaker, approximately 100 frames were used to generate both the mean value and standard deviation.

modified by passing the flat-spectrum excitation signal through an all-pole filter whose transfer function is described by Eq. (7). We tried both methods and noted virtually no difference in the sound quality.

Test and Evaluation

We incorporated the amplitude spectral modification of the voiced excitation signal in NRL's programmable real-time narrowband voice processor and in another voice processor currently under development. We used the Diagnostic Acceptability Measure (DAM) to evaluate the speech quality of these two systems. Both tests yielded virtually identical results. A 5-point improvement was shown in the overall DAM scores, indicating that the speech quality of our modified LPC is closer to that of the 9.6 kb/s APC than to the conventional 2.4 kb/s narrowband LPC (Fig. 4).

Though we did not expect the amplitude spectrum modification of the voiced excitation signal to noticeably affect consonant intelligibility, we nevertheless conducted Diagnostic Rhyme Tests (DRTs) to ensure that it did not hurt the speech intelligibility. The DRT scores for three male and three female speakers in a quiet environment were 87 both with and without the amplitude spectrum modification. Likewise, the DRT scores for three male speakers in a shipboard environment were virtually unchanged: 78 with modification and 77 without modification. These results confirm that our amplitude spectral modification of the voiced excitation signal significantly improves the quality of the narrowband LPC speech without affecting the intelligibility.

PHASE SPECTRUM SHAPING OF THE VOICED EXCITATION SIGNAL

Before there was a convenient way to generate complex signals with independently controlled phases it was thought that the human ear was phase deaf. Today we can adjust the phase spectrum of a complex waveform easily, and studies have found that the phase relationships between tones do have some influence on the perceived sound quality. For example, every experienced organist prefers the sound of an organ having individual oscillators (such as Conn, Allen and Rodger organs) over the sound of an organ with only 12 master oscillators that regenerate all the harmonically related tones (such as Baldwin or Hammond organs). Though difficult to describe, there is something more pleasing about complex waveforms with incoherent phases.

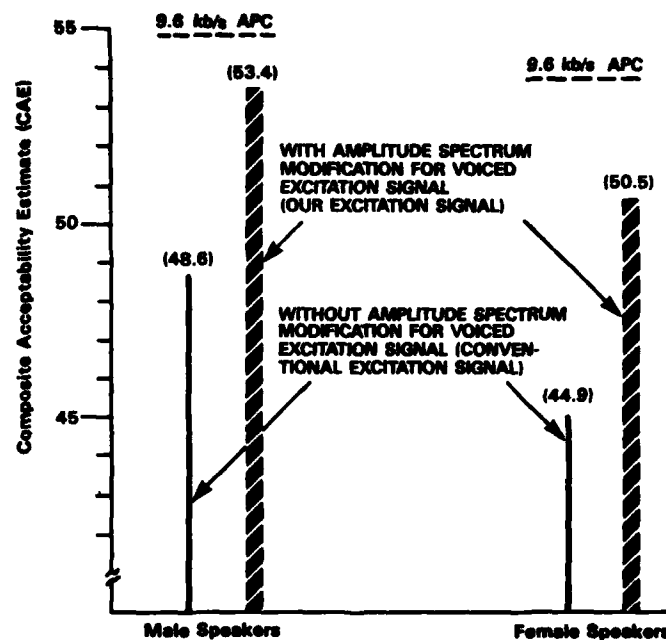


Fig. 4 — DAM scores for the 2.4 kb/s narrowband LPC. This figure illustrates the degree of improvement in the speech quality as a result of the amplitude spectral modification of the voiced excitation signal in the 2.4 kb/s LPC. For purposes of illustration, the DAM scores for the 9.6 kb/s APC voice processor are also shown.

Similarly, a number of practitioners in the speech analysis and synthesis fields have observed that the perceptual quality of synthetic speech depends to some extent on the phase spectrum of the voiced excitation signal [9]. Some have even observed that a reduction of peakiness in the voiced excitation signal, which is related to the phase spectrum, results in a reduction of buzziness in the synthetic speech [10]. In any case, the phase spectrum of the voiced excitation signal does not affect the pitch [11].

Ideally, the phase spectrum of the voiced excitation signal should be the phase spectrum of the pitch-synchronously windowed prediction residual with a window width equal to the pitch period. If both amplitude spectra are equal, the resulting excitation signal is equal to the prediction residual of one pitch period—the ideal excitation signal for a pitch-excited LPC or an LPC that repeats the voiced excitation signal at the pitch rate. Actually, some researchers have suggested using the median differential delay of the pitch-synchronously windowed prediction residual (defined as the first derivative of the phase spectrum with respect to frequency) [12,13] to determine the preferred phase spectrum of the excitation signal. The median delay is an approximately linearly ascending function of frequency, with a total increment of delay of roughly 1.2 ms from 0 Hz to the upper cutoff frequency of 3.2 kHz. The resulting sound quality is reported to be more natural than when a constant differential delay of zero (i.e., an impulse train) is used. As it turns out, the stationary part of the differential delay of our voiced excitation signal is quite similar to the median delay of the pitch-synchronously windowed prediction residual mentioned above. We use a time-dispersed voiced excitation signal for two reasons: (a) to improve the performance in tandem with continuously variable slope delta (CVSD) systems, and (b) to best use the available dynamic range of the (arithmetic) processor used.

The time-invariant portion of the phase spectrum discussed above fully specifies the conventional voiced excitation signal. The phase spectrum of our voiced excitation, however, has an additional time-variant portion to accommodate a small amount of waveform variation from one pitch cycle to the next. These period-to-period waveform variations, often referred to as pitch jitter, are caused in part by irregularities in vocal cord movement, and in part by the turbulent air flow from the lungs during the glottis-open period of each cycle. The amount of jitter varies with the fundamental pitch frequency, the age of the speaker, his or her nervous condition, and the degree of muscular elasticity.

Without an appropriate amount of pitch jitter, the synthetic speech sounds unnatural in several ways. First, it sounds flat and machinelike because the waveform is too similar from one pitch cycle to the next. Second, the synthetic speech sounds heavy and buzzy because of a lack of change, or flutter, particularly in the higher pitch harmonics. A combination of these characteristics makes the synthetic speech sound edgy and tense, though most people are only subconsciously aware of it.

This last effect deserves special attention because of its particularly insidious nature. When we look at the structure of a soothing, mellifluous voice like President Reagan's, we immediately notice that such a voice lacks the strong, regular pitch harmonics so prevalent in the synthetic LPC speech. We believe this is due to the presence of a certain amount of breath air during the glottis-open period, which introduces flutter in the high-frequency pitch harmonics. On the other hand, strong, regular pitch harmonics similar to those of the LPC synthesized speech are characteristic of sharp, clear voices like Paul Harvey's, and of speakers who are tense or angry. This is probably caused by a stiffening of the vocal cord muscles.

Figures 5 through 7 vividly illustrate how the speech and prediction residual waveforms differ in unusually mellow, normal, and tense voices for both male and female speakers. Note that the periodicity of the prediction residual, particularly that of the high-passed prediction residual, is progressively better defined as the tenseness of the voice increases. In very tense voices the prediction residual looks much like the conventional voiced excitation signal used in the narrowband LPC (see Fig. 8). We believe this is one of the reasons LPC speech sounds unnecessarily tense regardless of the quality of the speaker's voice.

All these observations lead us to the conclusion that a small amount of irregularity in the narrowband LPC speech is highly desirable. A similar conclusion was reached by Makhoul et al. [14], who introduced irregularity in LPC synthesized speech by using a mixed source in which the periodic pulse train was low-pass filtered while the noise was high-pass filtered at the same cutoff frequency. The cutoff frequency was variable and was estimated to be the highest frequency at which the speech spectrum was considered periodic. This cutoff frequency was quantized into 2 or 3 bits and transmitted to the receiver. The frequency quantization step was as coarse as 500 Hz, and low-order Butterworth filters were used. According to the authors, the above mixed excitation source appeared to reduce two seemingly different types of buzziness: the first was the quality of synthetic voiced fricatives; the second was the buzziness of sonorants, associated mainly with low-pitched voices.

Mixed excitation sources are not new; they have previously been applied to channel vocoders [15,16] and to the formant synthesizer [17] to improve voice quality. Our improvement to the LPC excitation signal also uses a mixed excitation source. In our approach, the mixed excitation source is simply a special case of the excitation signal generator described in Eq. (1) and can have both pitch-epoch variations and period-to-period waveform variations. Because we are constrained by the DoD interoperability requirements we cannot use any information not transmitted by the standard narrowband LPC. While some flexibility is lost by not using this additional information, our mixed excitation source is still much closer to the ideal excitation for the LPC analysis/synthesis system (i.e., the prediction residual) than is the conventional excitation.

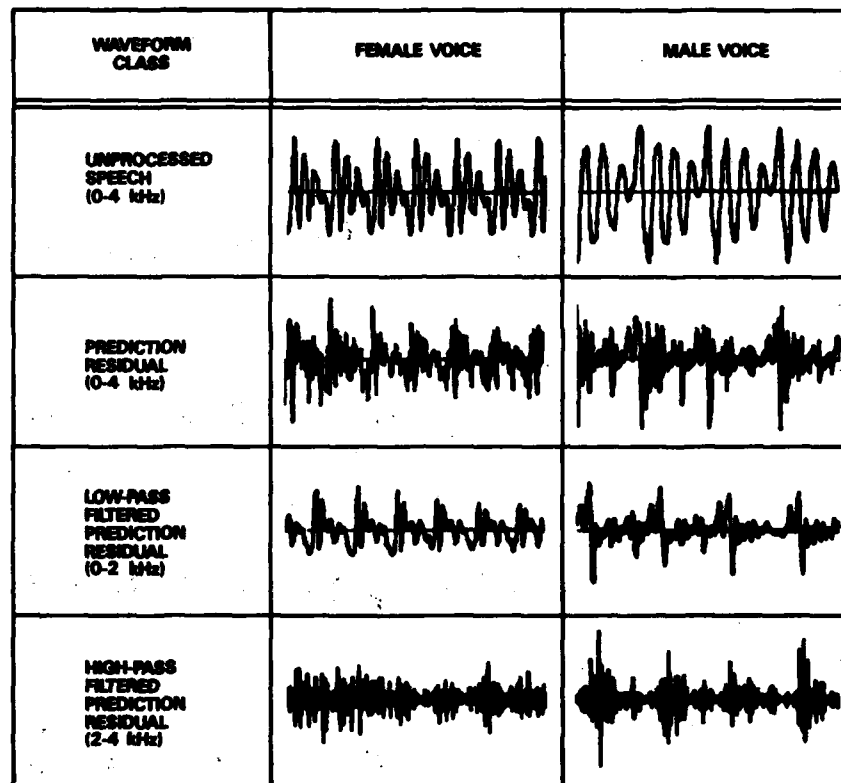


Fig. 5 — Unprocessed speech and prediction residual waveforms of *soothing, mellow* voices. Note the randomness of the prediction residual, particularly the high-passed prediction residual, and compare this waveform with the conventional narrowband LPC voiced excitation signal shown in Fig. 8. Some amount of randomness in the excitation signal is essential for the production of natural sounding speech. Note also the highly oscillatory speech waveform characteristic of mellow voices. The prediction residual waveforms illustrated in this figure (as well as those in Figs. 6 and 7) have been amplified four times for clarity.

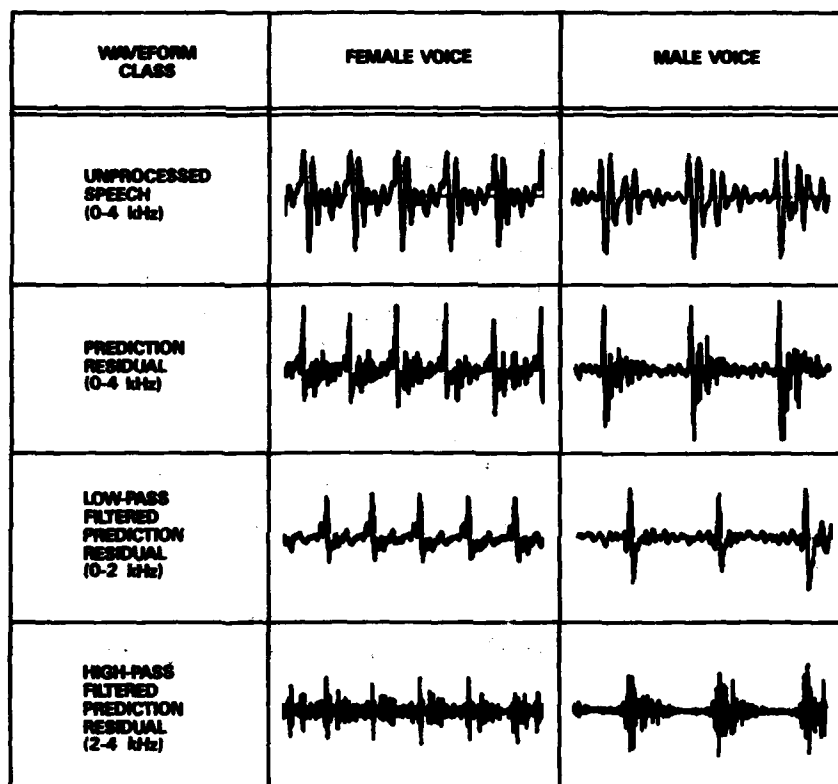


Fig. 6 — Unprocessed speech and prediction residual waveforms of *normal* voices. Note that the periodicity of the prediction residual is better defined than in the preceding figure, but less than for the tense voices in the following figure. Figure 8 illustrates that our voiced excitation signal for the narrowband LPC has a similar amount of randomness.

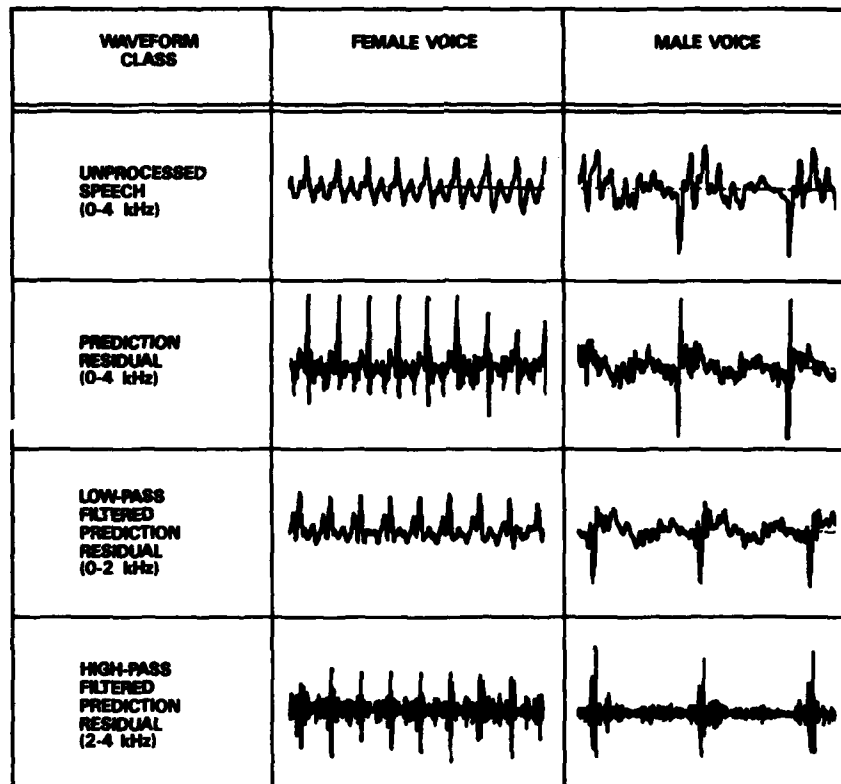


Fig. 7 — Unprocessed speech and prediction residual waveforms of *sense* voices. Note that the well-defined periodicity of the prediction residual, even the high-passed prediction residual, is very similar to that of the conventional narrowband LPC voiced excitation signal (Fig. 8). Note also the highly damped speech waveform which might easily be mistaken for a seismic wave.

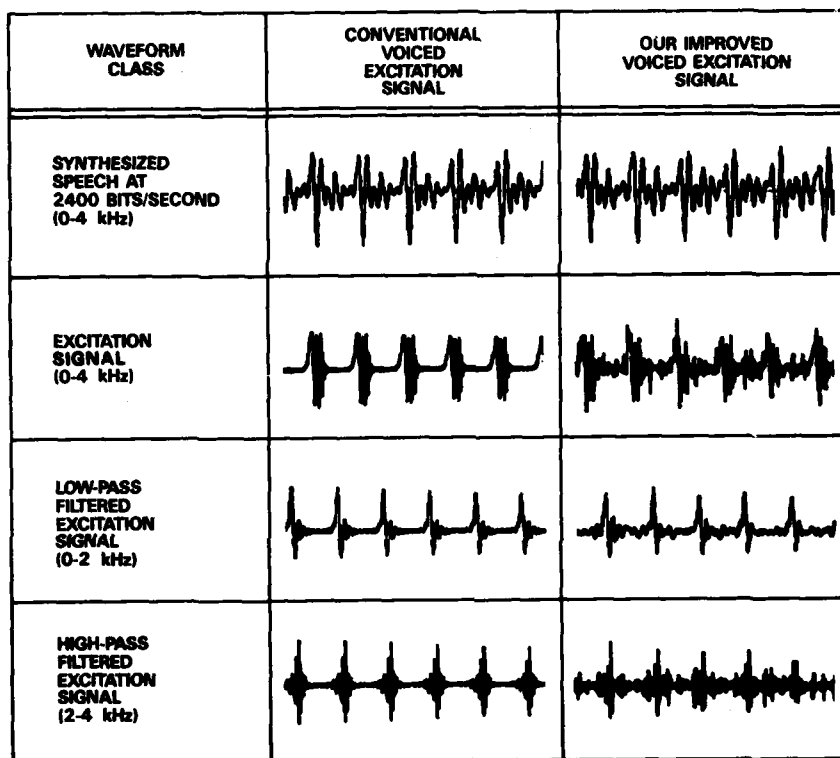


Fig. 8 — Synthesized speech and excitation signal waveforms for the narrowband LPC. These waveforms are generated by the use of LPC parameters extracted from the normal female speech waveform shown in Fig. 6. The absence of randomness in the conventional voiced excitation signal is in part responsible for the tense and unnatural speech quality of the narrowband LPC. (Compare the left column of this figure with Fig. 7.) The presence of randomness in our voiced excitation signal (right column) adds naturalness to the synthesized speech. Our voiced excitation signal is an approximation of the actual prediction residual of the normal female voice shown in Fig. 6.

The phase spectrum $\phi(k)$ of our excitation signal as expressed by Eq. (1) consists of two parts:

$$\phi(k) = \phi_0(k) + \Delta\phi(k) \quad k = 1, 2, \dots, K, \quad (9)$$

where $\phi(k)$ and $\Delta\phi(k)$ are the k th stationary and random phase components respectively. The random part of the phase spectrum is further divided into two parts:

$$\Delta\phi(k) = \Delta\phi_1(k) + \Delta\phi_2(k) \quad k = 1, 2, \dots, K, \quad (10)$$

where $\Delta\phi_1(k)$ and $\Delta\phi_2(k)$ are the random phases contributing to pitch-epoch jitter and period-to-period waveform variations respectively. We discuss these phase spectral components in the following section.

Stationary Part of the Phase Spectrum

The stationary part of the phase spectrum of the voiced excitation signal is important because it has a direct bearing on the peakedness and dispersiveness of the excitation signal. For example, if the phase spectrum is a linear function of frequency, or the differential delay is zero, all the frequency components will be phase-aligned and will produce a spike or impulse.

The use of an impulse for the voiced excitation is undesirable for two reasons. First, a spiky excitation signal produces a spiky narrowband LPC output which does not operate well in tandem with high-rate voice processors that encode the difference of two consecutive speech samples, such as continuously variable slope delta (CVSD) systems. Because CVSD cannot accurately follow the steep changes in the input amplitude produced by the impulse excitation, the output speech is distorted. Over the years, the narrowband LPC has improved its tandem performance with the CVSD. At one time the DRT score for a 16 kb/s CVSD operating from the narrowband LPC output was 78 for three male and three female voices; it is now 82. One of the major reasons for this improvement is the use in the LPC of a time-dispersed voiced excitation signal in lieu of an impulse excitation.

Second, a spiky excitation signal requires a greater dynamic range in the LPC signal processor, so the output amplitude often has to be lowered to avoid clipping. We can reduce the required dynamic range by as much as 10 dB by using a time-dispersed voiced excitation signal like that discussed below.

On the other hand, it is also undesirable for the voiced excitation signal to be dispersed over several pitch periods because the LPC synthesizer is a dynamic system in which the filter coefficients are updated pitch synchronously. The problem is even more complicated because the current narrowband LPC calibrates the speech level after the synthesis, with a constant power excitation at the input. For proper superposition and calibration, the output waveform generated by each set of excitation signal samples and filter coefficients must be stored independently. In general, a shorter excitation signal requires less data storage and fewer computations.

In the past, a number of different approaches have been investigated in an effort to design a family of signals with flat amplitude spectra and low peak amplitudes [9,18]. If the signal is expressed as a Fourier series, like our excitation signal, the required phase spectrum is a quadratic function of frequency [9].

Thus,

$$\phi_0(k) = (2\pi)\xi \left(\frac{k}{K} \right)^2 \quad k = 0, 1, \dots, K \quad (11)$$

where $\phi_0(k)$ is the k th stationary phase component defined in Eq. (1), K is the number of spectral components defined in Eq. (2), and the quantity ξ is an integer number—the larger the ξ , the greater the dispersion of the excitation signal. The differential delay, as obtained from Eq. (11), is

$$\begin{aligned}
D_0(k) &= \frac{\Delta\phi(k)}{\Delta\omega} \\
&= \frac{\left(\frac{1}{2}\right) [\phi(k) - \phi(k-1)]}{\Delta\omega} \\
&= \frac{(2\pi)(2\xi)}{K(\Delta\omega)} \left(\frac{k}{K}\right)
\end{aligned} \tag{12}$$

in which $\Delta\omega$ is a uniform frequency spacing between two adjacent spectral components. In our narrowband LPC, $K(\Delta\omega)$ is $(2\pi)4000$ rad/s. Thus, Eq. (12) may be written as

$$D_0(k) = \left(\frac{\xi}{2}\right) \left(\frac{k}{K}\right) \text{ ms.} \tag{13}$$

Equation (13) states that if the phase angle is a multiple of 2π rad at 4000 Hz, the differential delay at the same frequency is a multiple of 0.5 ms.

For purposes of illustration, we generated four different voiced excitation signals using $\xi = 3, 4, 5$, and 6 in Eqs. (11) and (13). Table 3 lists the spectral and temporal characteristics of these signals. In Example 1 ($\xi = 3$) the differential delay increases linearly from 0 ms at 0 Hz to 1.5 ms at 4000 Hz. Table 4 shows the excitation signal samples which are dispersed over 25 sampling time intervals. The peak amplitude reduction factor—defined as the maximum signal magnitude when the signal is normalized to have a unity power—is 8.98 dB. This is an impressive figure since the peak amplitude reduction factor realized by the 40-sample voiced excitation signal currently used by the DoD narrowband LPC is only 9.18 dB. In the second example ($\xi = 4$), the differential delay at 4000 Hz is increased to 2 ms, and the excitation signal samples are dispersed over 31 sampling time intervals. The resulting peak amplitude reduction factor is increased to 9.51 dB, and so on.

For our excitation signal we set $\xi = 3$ in Eqs. (11) and (13) (Example 1) because this yields a good peak amplitude reduction factor for the duration of the excitation signal. To verify that this 25-sample excitation signal can reproduce the originally specified frequency spectra characteristics, we computed both the amplitude and the phase spectra. (We feared that integerization and truncation of samples might have produced some spectral error.) Figure 9 shows that the computed spectra are virtually identical to the originally specified spectra.

Table 3—Characteristics of Stationary Part of Voiced Excitation Signals

| Example | Amplitude Spectrum | Phase Shift ^b @ 4000 Hz (2π) ξ (rad) | Diff. Delay ^c @ 4000 Hz 0.5 ξ (ms) | Absolute Maximum Amplitude When $\sum e^2(n) = 1$ (dB) | Dispersion Width ^d (No. of Samples) |
|----------------|--------------------|--|--|---|--|
| 1 ^a | Flat | 3(2π) | 1.5 | 0.3555 -8.98 | 25 |
| 2 | Flat | 4(2π) | 2.0 | 0.3344 -9.51 | 31 |
| 3 | Flat | 5(2π) | 2.5 | 0.3194 -9.91 | 35 |
| 4 | Flat | 6(2π) | 3.0 | 0.2835 -10.95 | 41 |

^aOur choice.

^bThe phase spectrum is a quadratic function of frequency.

^cThe differential delay is a linear function of frequency.

^dFor comparison purposes, the dispersion width is arbitrarily defined as the time interval in which every sample has a magnitude $\geq 1/256$ when the signal amplitude has normalized to have a unity power.

Table 4—Sample Values of the Stationary
Part of Voiced Excitation Signals

| Time Index | Example | | | | |
|---------------|----------------|------|------|------|--------|
| | 1 ^a | 2 | 3 | 4 | |
| 1 | | | | 4 | |
| 2 | | | | -6 | |
| 3 | | | | 8 | |
| 4 | | | -5 | -12 | |
| 5 | | | 7 | 19 | |
| 6 | | -4 | -11 | -29 | |
| 7 | | 6 | 17 | 44 | |
| 8 | | -10 | -28 | -69 | |
| 9 | 5 | 16 | 44 | 104 | |
| 10 | -8 | -26 | -72 | -154 | |
| 11 | 13 | 44 | 114 | 212 | |
| 12 | -24 | -76 | -175 | -262 | |
| 13 | 43 | 128 | 244 | 267 | |
| 14 | -81 | -204 | -295 | -183 | |
| 15 | 147 | 289 | 271 | -9 | |
| 16 | -252 | -335 | -113 | 228 | |
| 17 | 359 | 239 | -155 | -290 | |
| 18 | -364 | 44 | 327 | 60 | |
| 19 | 92 | -342 | -152 | 253 | |
| 20 | 336 | 231 | -245 | -194 | |
| 21 | -306 | 250 | 225 | -212 | Center |
| 22 | -336 | -231 | 245 | 194 | |
| 23 | 92 | -342 | -152 | 253 | |
| 24 | 364 | -44 | -327 | -60 | |
| 25 | 359 | 239 | -155 | -290 | |
| 26 | 252 | 335 | 113 | -228 | |
| 27 | 147 | 289 | 271 | -9 | |
| 28 | 81 | 204 | 296 | 183 | |
| 29 | 43 | 128 | 244 | 267 | |
| 30 | 24 | 76 | 174 | 262 | |
| 31 | 13 | 44 | 114 | 212 | |
| 32 | 8 | 26 | 72 | 154 | |
| 33 | 5 | 16 | 44 | 104 | |
| 34 | | 10 | 28 | 69 | |
| 35 | | 6 | 17 | 44 | |
| 36 | | 4 | 11 | 29 | |
| 37 | | | 7 | 19 | |
| 38 | | | 5 | 12 | |
| 39 | | | | 8 | |
| 40 | | | | 6 | |
| 41 | | | | 4 | |

^aOur choice

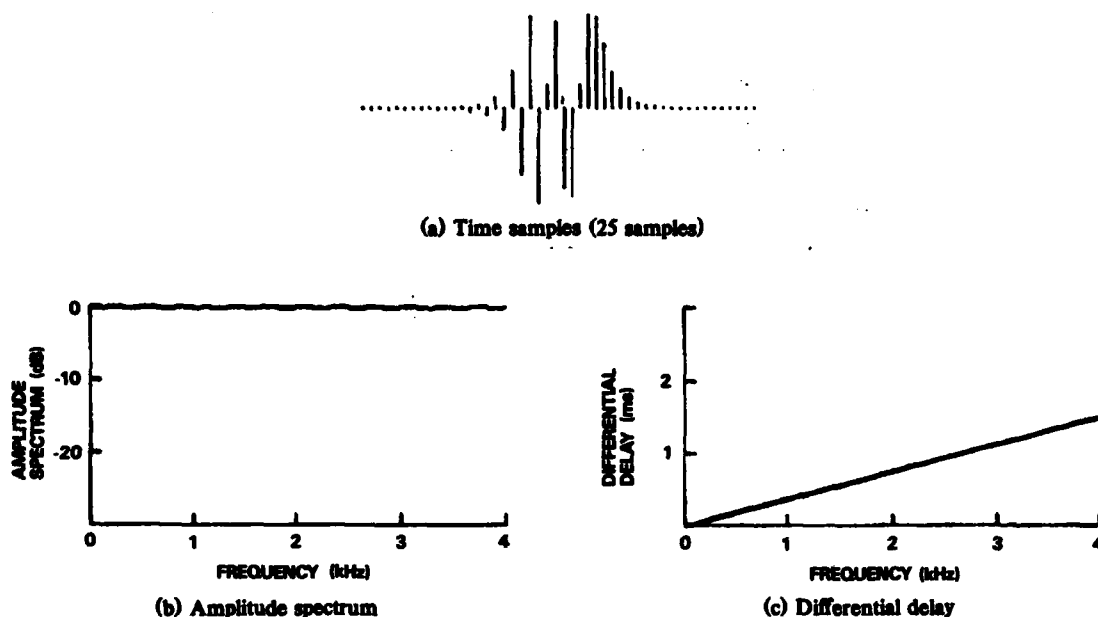


Fig. 9 — Our chosen stationary voiced excitation signal: time samples, computed amplitude spectrum, and differential delay. This is Example 1 in Tables 3 and 4 and is obtained by letting $\xi = 3$ in Eq. (11) or Eq. (13).

It is interesting to note that the delay shown in Fig. 9(c) is similar to the median delay computed from the actual prediction residual by Atal and David [13]. The median delay also increases nearly linearly with the increase in frequency. The total delay increment from 0 Hz to the highest frequency is approximately 1.2 ms, which is close to that shown in Fig. 9(c).

Random Part of the Phase Spectrum

As stated previously, there are two types of randomness present in the natural voiced speech waveform. One is pitch-epoch variation, or jitter, caused by irregularities in vocal cord movement; the other is period-to-period waveform variation caused by the turbulent air flow from the lungs. To incorporate these variations in the excitation signal we need two different kinds of random spectral components as discussed below.

Pitch-Epoch Variations

The magnitude of pitch-epoch variations is not large—the average shift is reportedly somewhere between 10 and 60 μ s for adult male speakers [19]. The presence of this small amount of pitch variation is nevertheless essential to make synthesized speech sound more natural. Because the pitch period as transmitted by the narrowband LPC is merely the average pitch period updated at a fixed frame rate (approximately two pitch periods for an average male speaker, and four pitch periods for an average female speaker), it does not contain any information related to pitch-epoch variation. Even if the pitch period were updated several times per frame, it still would not reflect the actual pitch-epoch variation because the pitch tracker has too much inertia to be influenced by such small changes. Moreover, the pitch period quantization, where the minimum pitch period resolution is one sampling time interval, or 125 μ s, is far too coarse to capture pitch-epoch variations as small as 10 to 60 μ s. In short, pitch-epoch variation in the narrowband LPC must be artificially introduced at the receiver.

In our voiced excitation signal, the pitch epoch is readily altered by allowing an additional linear phase in the phase spectrum as expressed by Eq. (1). The gradient of the linear phase is randomly perturbed from one pitch period to the next. As an example, if the phase changes linearly from 0 rad at

0 Hz to 1 rad at 4 kHz, the resulting differential delay of the time waveform is $1/8000\pi$ second or $39.789 \mu\text{s}$. A smaller phase shift gives rise to a proportionally smaller shift in pitch epoch. We found a maximum jitter of $10 \mu\text{s}$ to be satisfactory. Thus the phase shift at 4 kHz is a maximum of $1/4$ rad and is computed by

$$\Delta\phi_1(k) = \frac{m}{4} \left[\frac{k}{K} \right] \text{ rad} \quad k = 1, 2, \dots, K \quad (14)$$

where $\Delta\phi_1(k)$ is the random part of the phase spectrum contributing to pitch epoch variations, k is the frequency index, K is the total number of frequency components, and m is a uniformly distributed random number between 1 and -1 which changes at each pitch epoch.

It is worth noting that even under the most ideal operating conditions (such as noise-free speech and error-free transmission) the narrowband LPC generates a considerable amount of pitch irregularity, or flutter, in the synthesized speech. This is primarily because the LPC analysis window is not placed in perfect synchrony with the pitch cycle. This effect is further aggravated by the parameter quantization, which tends to cause the synthesized speech waveform to vary even when the input is well sustained. Since the narrowband LPC updates the speech parameters once every frame, the frequency of the flutter is fairly low, and our ears are rather sensitive to it. Therefore, the pitch-epoch jitter must not reinforce the already audible low-frequency flutter. (Note that flutter of this kind would not exist in a speech synthesis system where the speech data are defined at irregular and sparsely spaced time intervals. However, in this case the magnitude of the minimum pitch-epoch jitter would be even greater than that of the narrowband LPC.)

Period-To-Period Waveform Variations

The period-to-period waveform variations caused by breath air are very complex. On the one hand they are random because the air coming from the lungs is turbulent. On the other hand they are pitch-modulated because the air passes through the glottis as it opens and closes at the pitch rate. The period-to-period waveform variations in the prediction residual (the ideal excitation signal) are disproportionately strong in the high-frequency regions because the LPC analysis filter boosts the treble to flatten the spectral envelope of the speech, but not that of the breath noise. Figures 5 through 7 show that the amount of period-to-period waveform variation in the prediction residual differs substantially from speaker to speaker. In addition, evidence indicates that the amount of waveform variation depends on the speech sound; for example, there is more randomness in back vowels than in front vowels.

Period-to-period waveform variations are caused by a multitude of factors that cannot be emulated by a simple mixed excitation source, nor by our general form of the mixed excitation source, when relevant information is not available at the receiver. Because a many-to-one transformation exists between random noise and its perception by the human ear, the nature of any artificially introduced randomness in the voiced excitation signal need not be exactly identical to that of the prediction residual. For example, unvoiced sounds from the telephone are severely distorted, yet we can still identify them. Similarly, the spectral distribution of a fricative sound varies widely from speaker to speaker [20], but this does not cause any misunderstanding. According to a recent experiment at NRL, the intelligibility of the narrowband LPC speech is virtually unaffected even when the set of LPC coefficients from unvoiced speech is quantized very coarsely into an eight-bit quantity (i.e., one of only 256 possible combinations).

We listened to a large number of speech samples processed by our real-time narrowband LPC as we varied the nature of the random components in the voiced excitation signal. While there seemed to be a wide range of acceptable characteristics, we noted that the overall intensity and the frequency distribution of the random components appeared to be more significant than other parameters. The

overall intensity is important because the speech quality suffers both if it is too low or if it is too high. The frequency distribution characteristics are also important because the speech sounds warbly if there is too much low-frequency jitter. Note that these are the only two parameters used by the narrowband LPC to synthesize unvoiced speech.

Unfortunately we cannot extract nor transmit these two parameters at the LPC transmitter because the resulting LPC would not be compatible with the standard DoD format. Therefore we would like to extract average values for these two parameters from the actual prediction residual so that we may use them as constants in the LPC receiver. This analysis is by no means straightforward; the selection of the proper prediction residual samples and the choice of the analysis method are both critical.

The prediction residual samples must be selected carefully because period-to-period waveform variations in the prediction residual are caused not only by breath noise and the instability of the excitation source (i.e., the glottis), but also by the changes in the vocal tract during speech transitions. Since we would like to exclude the effects of the speech transitions in the estimated parameters, we must select prediction residual samples from voiced frames where the LPC coefficients (i.e., the vocal tract filtering characteristics) do not vary significantly from one frame to the next. In other words, we must select the prediction residuals for analysis from sustained vowels.

Once the residual samples are selected, the choice of the analysis method is critical for obtaining reliable analysis results. The most direct way of estimating the intensity and frequency distribution parameters is through a variance analysis of the phase spectra derived from the prediction residual using a pitch-synchronous analysis window. However, we find this approach insurmountably difficult and risky since even visual inspection cannot reliably determine the pitch epoch from a highly noise-like prediction residual (for example, see Fig. 5). The phase spectrum is sensitive to the location of the window with respect to the waveform under analysis, and frequent window placement errors will degrade the estimated parameters beyond any usefulness. Since we are basically interested in the gross characteristics of the frequency dependency and the overall intensity, rather than their detailed frame-by-frame characteristics, we choose to use an alternate method of analysis.

This alternate method involves the spectral analysis of the pitch-filtered prediction residual defined by

$$r'(i) = r(i) - \beta r(i - T) \quad (15)$$

where $r(i)$ is a prediction residual sample, $r'(i)$ is a pitch-filtered prediction residual sample, T is the pitch period, and β is a first-order prediction coefficient of $r(i)$ T samples apart. As usual, β is obtained by minimizing the mean-square value of the right-hand member of Eq. (15). Thus,

$$\beta = \frac{\sum_i r(i)r(i - T)}{\sum_i r^2(i - T)} \quad (16)$$

Since we select only stationary prediction residuals for the analysis, β may be expressed by

$$\beta = \frac{\sum_i r(i)r(i - T)}{\frac{1}{2} \left[\left[\sum_i r^2(i) \right] + \left[\sum_i r^2(i - T) \right] \right]} \quad (17)$$

where the magnitude is bounded between 1 and -1 . Equation (15) represents the input-output relationship of a notch filter which suppresses harmonically related frequencies (in this case, the fundamental pitch frequency and its harmonics). The quantity β is related to the notch filter bandwidth and is

dependent on the randomness of the input. For example, in the absence of randomness, as in the conventional voiced excitation signal, β is unity. For actual prediction residuals from steady vowels, β lies somewhere between 0.7 and 0.9.

With a steady vowel as the input, the pitch-filtered prediction residual is mainly period-to-period waveform variations of the prediction residual. Thus, the spectral analysis of the pitch-filtered prediction residual indicates both the nature of the frequency dependency and the overall intensity of the random parts of the prediction residual. Figure 10 shows the amplitude spectra of pitch-filtered prediction residuals generated from the three types of female voice waveforms previously illustrated in Figs. 5 through 7. For reference, Fig. 10 also shows the amplitude spectra of the corresponding prediction residuals. Note that the irregular spectral pattern of the prediction residual (mainly in the high-frequency region) may or may not be related to the presence of period-to-period waveform variations. This irregularity may also be due to the relatively constant absorption of selected frequencies by the vocal tract.

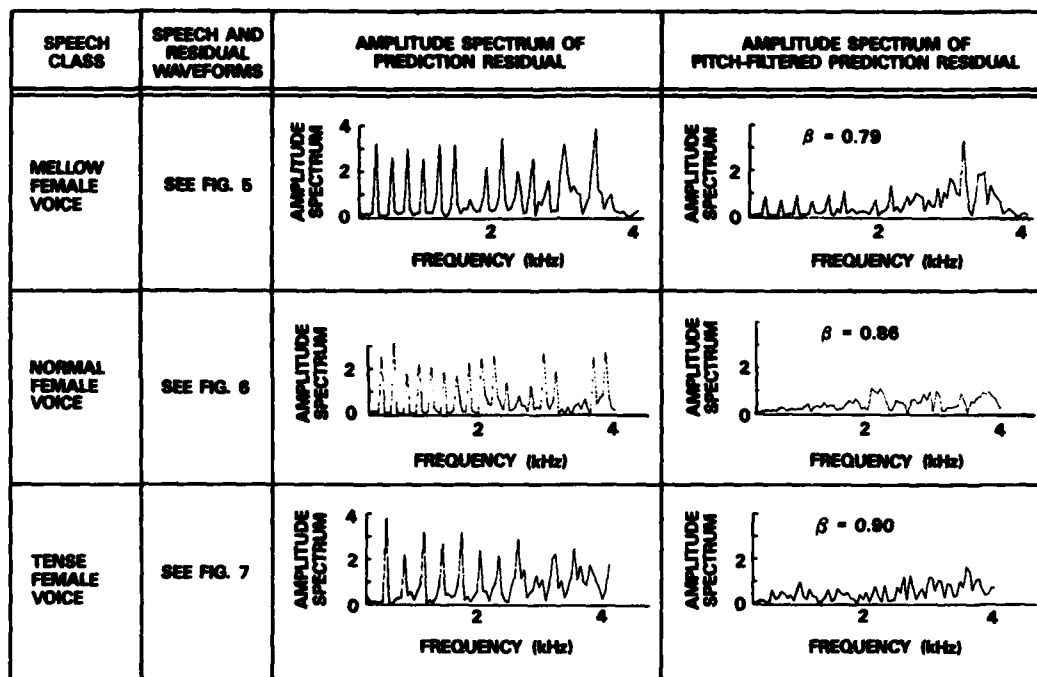


Fig. 10 — Amplitude spectra of prediction residuals and pitch-filtered prediction residuals from the three female voices shown in Figs. 5 through 7. As noted, the amplitude spectrum of the pitch-filtered prediction residual generally increases with frequency.

The spectral distribution of the pitch-filtered prediction residual is significant because it represents the spectrum of the period-to-period waveform variations in the prediction residual. We introduce random components in the voiced excitation signal such that the amplitude spectrum of the pitch-filtered excitation signal has a spectral distribution similar to that of normal voices as shown in Fig. 10. This figure as well as similar plots of other voices show that the amplitude spectrum of the pitch-filtered prediction residual is an approximately linear function of frequency, and the pitch prediction coefficient β is approximately 0.85. Thus the random part of the phase spectrum $\Delta\phi_2(k)$ is obtained numerically by using Eqs. (1), (15), and (17):

$$\Delta\phi_2(k) = \frac{\pi}{2}\sigma(k)\left[\frac{k}{K}\right] \text{ rad} \quad (18)$$

where $\sigma(k)$ is a uniformly distributed random variable between -1 and 1 , k is the frequency index, and K is the total number of components within the 0 to 4 kHz passband. Figure 11, which is similar to Fig. 10, compares the conventional voiced excitation signal and our modified voiced excitation signal. Note that our pitch-filtered excitation signal has characteristics more similar to those of the prediction residual of the normal voice. (The time samples of both excitation signals are shown in Fig. 8.)

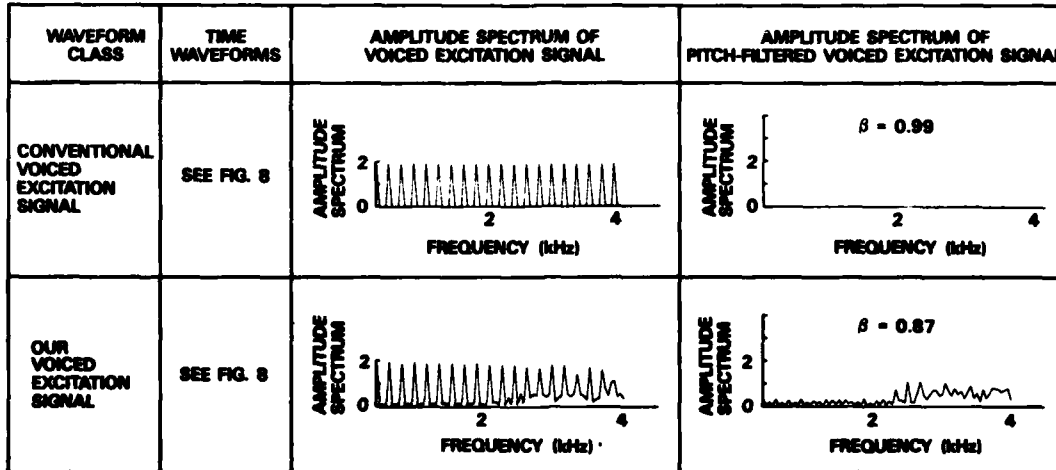


Fig. 11 — Amplitude spectra of the voiced excitation signal and the pitch-filtered voiced excitation signal for the conventional excitation (upper illustrations) and our modified excitation (lower illustrations). Both are derived from LPC parameters generated by using the speech waveform of the normal female voice shown in Fig. 6. (The prediction residual spectrum and pitch-filtered residual spectrum of this voice are shown in Fig. 10.) The conventional voiced excitation signal has a small amount of randomness because we carefully introduced the actual LPC parameter quantization and interpolation effects in the excitation signal, but the amount of randomness is negligible. On the other hand, our voiced excitation signal has randomness in which the frequency dependency and magnitude (in terms of the β value) are similar to those of the pitch-filtered prediction residual of the actual speech as shown in Fig. 10.

Test and Evaluation

When our voiced excitation signal is used in the narrowband LPC, one can readily hear that the output speech has a quality of breathiness not unlike that of the unprocessed speech. The output speech sounds much livelier, and the buzzy, twangy qualities often present in the conventional narrowband LPC output are greatly reduced. DAM tests were conducted to ascertain the degree of quality improvement achieved. The test results show a 4.7-point improvement for male speakers (from 48.6 to 54.3) and a 5.0-point improvement for female speakers (from 44.7 to 49.7). The scores for the modified LPC compare favorably with those for a 9.6 kb/s voice processor (54.8 for males and 53.5 for females). A DRT was also conducted to ensure that the phase spectral modification did not produce such strong improvements in speech quality at the expense of speech intelligibility. As expected, the DRT score of 85.8 for the modified LPC was only slightly better than the score of 85.3 for the conventional LPC.

MODIFIED UNVOICED EXCITATION SIGNAL

In the past, the unvoiced excitation signal has not received as much attention as the voiced excitation signal. The excitation signal traditionally used for generating all unvoiced sounds is simple random noise; no distinction is made between fricative sounds (/h/, /s/, /sh/, /f/, /th/) and burst, or stop, sounds (/p/, /t/, /k/). Usually the excitation signal is generated by randomly picking numbers from a table containing uniformly distributed random numbers; a small table containing about 256 numbers is adequate.

In our modified excitation signal generator both the voiced and unvoiced excitation signals are synthesized from Eq. (1). They only differ in their phase spectra: for the unvoiced excitation the phase spectral components are random variables, and may be distributed uniformly between $-\pi$ and π radians. According to the Central Limit Theorem our unvoiced excitation signal will actually tend to have a Gaussian distribution because each sample is expressed by a sum of random variables (Eq. 1). Figure 12 illustrates the probability density function of our excitation signal computed from 1000 samples having uniformly distributed phase spectral components. Figure 13 shows that the probability density function of our unvoiced excitation is approximately Gaussian, and it is actually a better approximation of the probability density function of the prediction residual of voiceless fricative speech than is the uniformly distributed unvoiced excitation signal used in the conventional narrowband LPC.

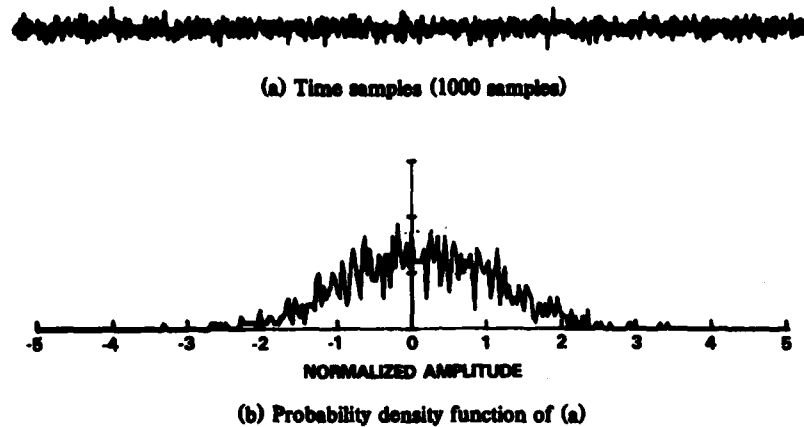


Fig. 12 — Characteristics of our unvoiced excitation signal used to generate the fricative sound /s/. The normalized amplitude is the excitation signal amplitude divided by its RMS value.

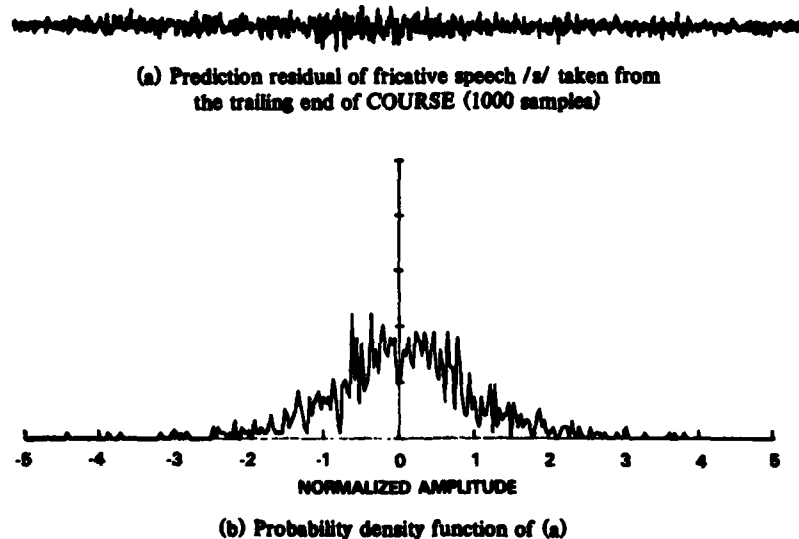


Fig. 13 — Prediction residual from an actual /s/. The probability density function shown here is similar to that of our unvoiced excitation signal for generating /s/ (Fig. 12). Note that the conventional unvoiced excitation signal is uniformly distributed noise.

Despite its inaccurate probability density function, the conventional unvoiced excitation signal is adequate for generating fricative sounds. This signal, the resulting synthesized speech waveforms, and the prediction residuals from such speech waveforms are basically stationary noise. Thus the ear tends to accept them as fricative sounds. However, this excitation is not satisfactory for generating burst sounds. The onsets of these sounds generate large spikes in the prediction residuals (Fig. 14), but the excitation signal conventionally used to synthesize them is still stationary noise. As a result CAT is often heard as HAT, and TICK may sound like THICK or SICK. To improve the reproduction of unvoiced bursts, we have modified the unvoiced excitation signal to include a way of generating such spikes.

This modified excitation signal is actually a superposition of two signals: one is similar to the conventional unvoiced excitation signal; the other is a train of randomly spaced pulses. The amount of pulse energy is proportional to the abruptness of the unvoiced speech as measured by the speech root-mean-square (RMS) ratio of two adjacent unvoiced frames. In the remaining part of this section we examine prediction residuals from both fricatives and abrupt unvoiced samples and compute the speech RMS ratios from various unvoiced onsets. We also present evidence demonstrating that the modified unvoiced excitation signal enhances the reproduction of unvoiced stops in the narrowband LPC.

Fricative Sounds and Their Prediction Residuals

In speech, fricative noise is generated by a turbulence in the airflow caused by a constriction somewhere in the vocal tract. The place of the constriction determines the frequency spectrum and the intensity of the sound. Figure 13 shows the amplitude distribution of the prediction residual processed from 1000 samples of /s/ at the trailing end of COURSE (female speaker). The amplitude distributions of the prediction residuals for other fricative sounds are similar to the example shown [20,21]. These distributions may be approximated by the Gaussian distribution function, and as such, the conventional excitation signal is adequate for producing these fricatives within the 4 kHz passband.

Unvoiced Plosives and Their Prediction Residuals

A plosive burst is a sequence of events that involves the integration of both spectral and temporal cues. First, a rapid closure is affected at some point in the oral cavity and pressure is built up behind it. When the closure is released a burst of energy having a broad bandwidth and short duration is generated. Unvoiced bursts (/p/, /t/, /k/) are louder and longer than voiced bursts (/b/, /d/, /g/) since more pressure is developed before release [21].

Because of this sudden burst of energy, the amplitude of the prediction residual of an unvoiced burst is particularly large at the onset of the sound. Therefore the accurate synthesis of unvoiced plosives requires an excitation signal having one or more sharp spikes at the onset. However, spikes should not be present at the onsets of fricative sounds. The implementation of such an excitation signal therefore requires a way of measuring the abruptness of the speech to discriminate between the burst onsets of stops and the relatively gentle onsets of fricatives. Because data rate restrictions prohibit the transmission of any additional information, this measure must be derived from the LPC parameters available at the receiver.

Measure of Abruptness

The abruptness of the speech is related to the amount of change in the speech energy over a short period of time. Thus the ratio of the speech RMS values from two consecutive frames should indicate the degree of abruptness. To test this hypothesis, we randomly selected words containing abrupt and nonabrupt unvoiced consonants and computed the speech RMS ratios at the consonant onsets. The test words were excerpted from casually spoken sentences, so they were not articulated any more carefully than would be expected in normal conversational speech. The computed speech RMS ratios, listed in

Table 5, are consistently larger for the stops and smaller for the fricatives. This is also true for the two words (TOOK and TOWN) contaminated by helicopter carrier noise.

Table 5—Speech RMS Ratios From Two Consecutive Unvoiced Frames

| Test Words (The underline indicates where the RMS ratio is computed) | | Ratio of Speech RMS Values from Two Consecutive Unvoiced Frames ^a |
|---|----------------------------|--|
| Abrupt Unvoiced Plosives | ou <u>t</u> | 14 |
| | st <u>o</u> p | 17 |
| | to | 32 |
| | blun <u>t</u> | 34 |
| | ca <u>n</u> | 19 |
| | ta <u>k</u> e | 20 |
| | co <u>u</u> rs <u>e</u> | 25 |
| | to <u>o</u> k ^b | 26 |
| | to <u>u</u> n ^b | 19 |
| Nonabrupt Unvoiced Fricatives | a <u>t</u> your | 22 |
| | pi <u>p</u> e | 11 |
| | st <u>o</u> p | 2 |
| | ge <u>l</u> f | 5 |
| | h <u>e</u> | 4 |
| | hi <u>s</u> | 3 |
| | sh <u>a</u> r <u>p</u> | 2 |
| | Fr <u>e</u> d | 2 |

^aRMS ratios less than 4 are set to 4 to reduce the effect of noise interference (see the text).

^bWith shipboard background noise

In general the presence of background noise decreases the magnitude of the speech RMS ratio, so unvoiced stops tend to sound like fricatives unless the noise interference is reduced somehow. For this reason we recommend the use of a noise-cancellation microphone and noise-suppression preprocessing, such as the spectral subtraction method [1], in noisy platforms. Table 6 lists the cumulative probability functions of background noise RMS values from eight different platforms by using both a noise-cancellation microphone and noise-suppression preprocessing. If the noise floor is less than 10 dB when the speech amplitude is quantized to 12 bits per sample, the effect of the noise floor on the RMS ratio is not significant. However, we set the minimum RMS at 4 in order to reduce the contrast between noise-free and noisy cases when computing the RMS ratio. The values in Table 5 were obtained on this basis.

Modified Unvoiced Excitation Signal Model

Our objective here is to improve the sound quality of unvoiced stops in the narrowband LPC by using only the information available at the receiver. We concluded that the best way to accomplish this was to modify the excitation signal by introducing sharp spikes as discussed above. In essence our modified unvoiced excitation signal is the conventional unvoiced excitation signal with a superimposed train of randomly spaced pulses. Thus, it may be expressed by

$$e(i) = n(i) + Rp(i) \quad (19)$$

Table 6—Cumulative Probabilities of Background Noise Amplitudes Observed at Eight Different Military Platforms

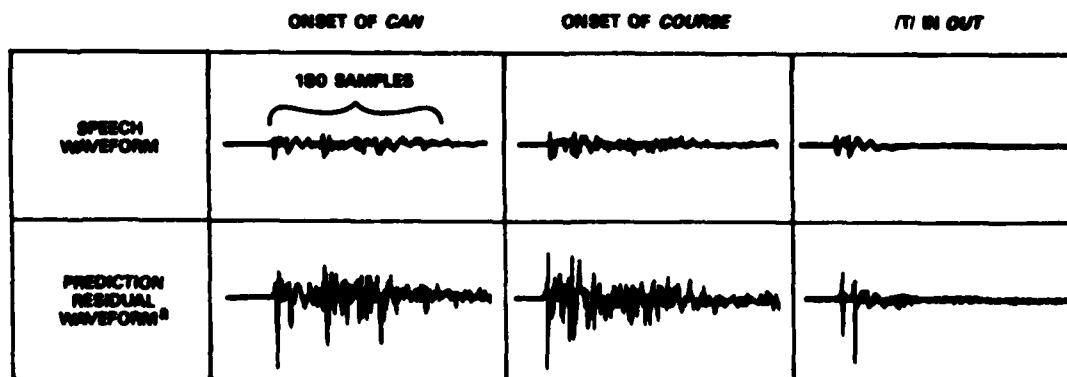
| Test Conditioning | Noise Level (dB) ^a | Narrowband LPC Amplitude Parameter ^b | | | | | | | | | | |
|-----------------------------|-------------------------------|---|------|------|------|------|------|------|------|------|------|------|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Quiet | — | 0.98 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Airborne command post noise | 85 | 0.76 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Shipboard noises | 82 | 0.91 | 0.94 | 0.96 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Office noise | 63 | 0.59 | 0.83 | 0.91 | 0.95 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 1.00 |
| E3A noise | 87 | 0.02 | 0.02 | 0.07 | 0.30 | 0.61 | 0.80 | 0.90 | 0.96 | 0.98 | 1.00 | 1.00 |
| Helicopter carrier noise | 76 | 0.19 | 0.71 | 0.90 | 0.95 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| P3C turboprop noise | 105 | 0.01 | 0.01 | 0.01 | 0.14 | 0.52 | 0.83 | 0.93 | 0.97 | 0.98 | 0.99 | 1.00 |
| Jeep noise | 92 | 0.02 | 0.05 | 0.16 | 0.49 | 0.77 | 0.88 | 0.94 | 0.97 | 0.99 | 1.00 | 1.00 |
| Tank noise | 112 | 0.02 | 0.03 | 0.14 | 0.39 | 0.66 | 0.82 | 0.89 | 0.93 | 0.96 | 0.99 | 1.00 |

^aThe normal speaking level is approximately 110 dB sound pressure level (SPL) at the microphone located 6 mm (1/4 inch) away from the mouth.

^bThe narrowband LPC amplitude parameter is the root-mean-square value of the preemphasized speech waveform. It is expressed in an integer number between 0 and 512.

where $e(i)$ is the modified unvoiced excitation signal, $n(i)$ is the conventional unvoiced excitation signal having one unit of RMS value, and $p(i)$ is the pulse train yet to be discussed. The quantity R , a factor proportional to the speech RMS ratio discussed in the preceding section, is updated at each frame. Note that the superposition of a pulse train onto the conventional excitation signal does not make the synthesized speech any louder, even if R is greater than zero, because the synthesized speech amplitude is calibrated by the same speech RMS value regardless of the nature of the excitation signal used.

The random spike component of the modified unvoiced excitation signal is dominant only at the onsets of unvoiced stops, and then usually for a single isolated frame (Fig. 14). Since the human ear cannot accurately analyze the turbulent speech waveform over such a short period of time, the exact nature and location of the spikes is not terribly critical. After examining numerous residual samples from unvoiced stops and conducting listening tests with synthesized stops, we decided to use four randomly spaced spikes per frame (Fig. 15).



^aamplified 4 times for larger display

Fig. 14 — Three examples of unvoiced plosives and their prediction residuals. Note large spikes in the prediction residual at the onsets. Without those spikes, the plosives often sound more like fricatives.

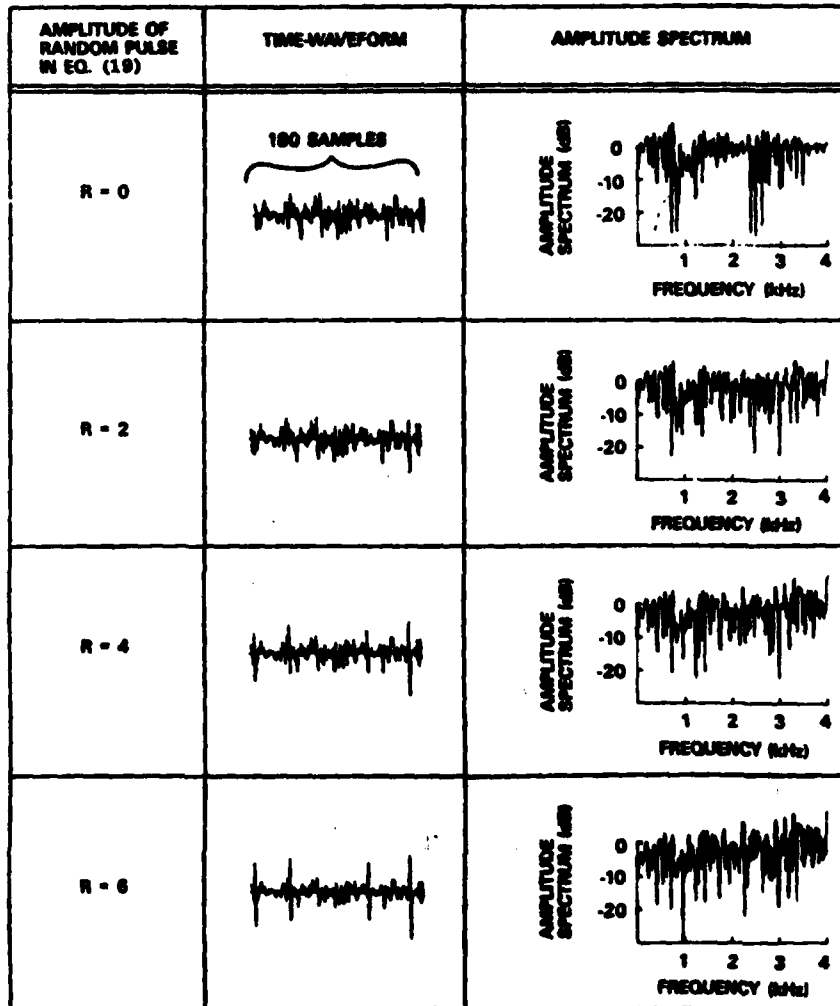


Fig. 15 — Our unvoiced excitation signals and their amplitude spectra. The presence of spikes in our unvoiced excitation signal improves the production of plosives. The quantity R is related to the speech RMS ratio across two adjacent unvoiced frames. When R is zero, the resulting waveform is the conventional unvoiced excitation signal. The amplitude spectrum of our unvoiced excitation signal does not show any undesirable resonant frequencies.

We observed that the greater the jump in speech RMS between two adjacent unvoiced frames, the greater the amplitude of the prediction residual spikes. Therefore we made the amplitude of each pulse, denoted by R in Eq. (19), proportional to the speech RMS ratio. As defined previously, $R = 1$ implies that each pulse amplitude is equal to the RMS value of the random component $n(i)$ in Eq. (19). Figure 15 shows that when $R = 6$ the resulting spike amplitude is sufficient for even the most distinctive stop bursts whose RMS ratios are around 25 (see Table 5). Therefore a reasonable value for R is

$$R = (\text{Speech RMS Ratio})/4 \quad (20)$$

where R is limited to a minimum of 0 and a maximum of 6. The pulses are spaced randomly so that they do not introduce harmonically related frequencies similar to pitch or formant frequencies.

The strong unvoiced plosive bursts produced by our modified unvoiced excitation signal can easily be seen in Fig. 16(b). When compared to the output of the conventional LPC (Fig. 16(c)) it is clear that the burst information present in the original speech (Fig. 16(a)) has been reproduced much more accurately by our unvoiced excitation signal. This results in clean, sharp plosive onsets and improves the intelligibility of these sounds noticeably—COURSE no longer sounds like HORSE, nor PEN like HEN.

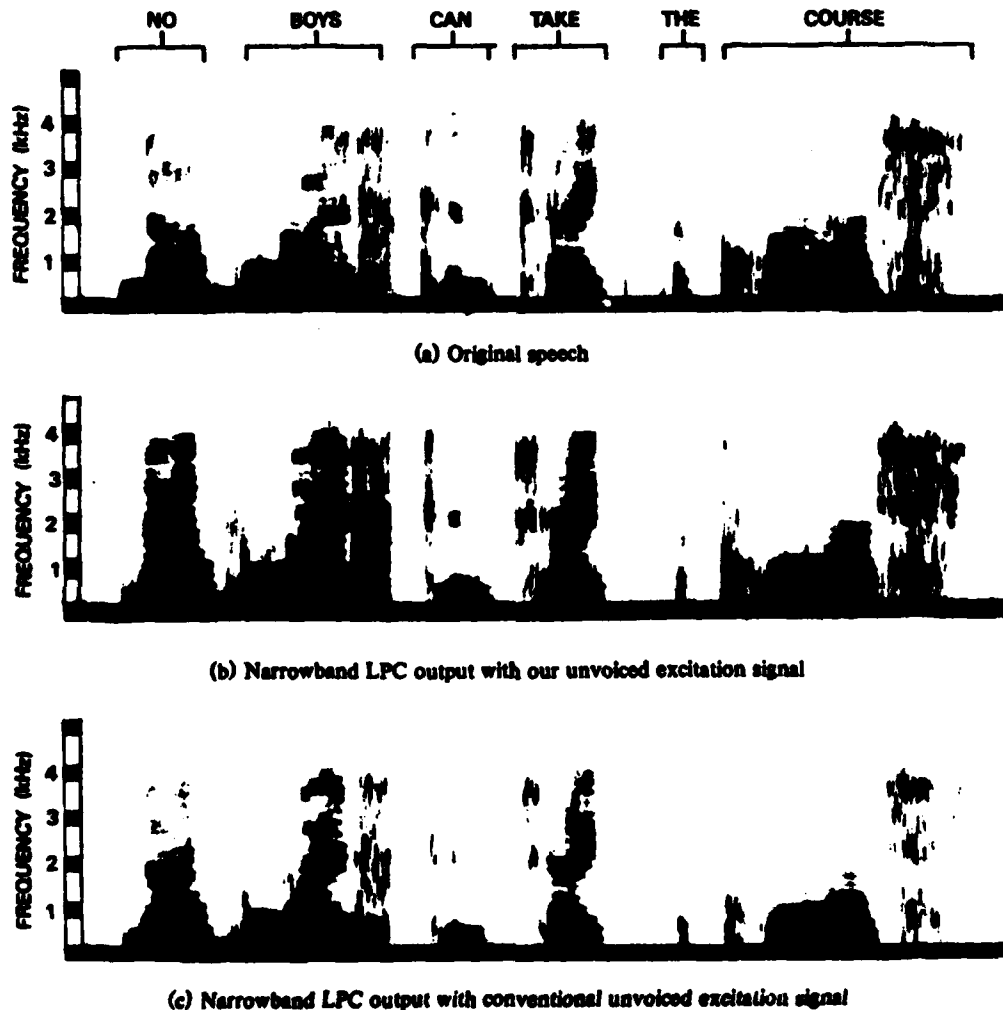


Fig. 16 — Spectrograms of narrowband LPC input and output. When our unvoiced excitation is used, the onsets of CAN, TAKE, and COURSE are reproduced better at the narrowband LPC output. Note the sudden bursts of speech energy at these onsets in Fig. 16(b) and compare them with those in Fig. 16(c).

Test and Evaluation

Our modified unvoiced excitation signal was developed to improve reproduction of unvoiced speech, in particular unvoiced plosives. The DRT is an excellent means for evaluating this improvement because it specifically tests the intelligibility of initial consonants including unvoiced plosives. We selected female speakers for the testing because the performance of the narrowband LPC is notoriously poorer with female voices than with male voices (average DRT scores are about 5.5 points lower).

Table 7 lists DRT scores for three female speakers using the narrowband LPC with the conventional unvoiced excitation signal and with our modified unvoiced excitation signal. The improvement for the attribute "graveness" is highly significant. A look at the score changes for the features within graveness reveals that this improvement is due primarily to better reproduction of unvoiced sounds, particularly plosives.

Table 8 lists the four features within graveness and the test words associated with each feature. When the attribute graveness is present, the loci of the second and third formants are relatively low; when this attribute is absent, they are relatively high. In both cases our unvoiced excitation signal produces higher scores for all sounds, particularly unvoiced plosives.

Table 7—DRT scores of narrowband LPC-processed speech for three females. The first set of scores was obtained using the conventional unvoiced excitation signal; the second set was obtained using our unvoiced excitation signal. Note the significant difference in the score for graveness which tests /p/ vs /t/, /f/ vs /t/, among others.

| Sound Class | Score | | |
|-------------|--|-------------------------------------|--------|
| | With Conventional Unvoiced Excitation Signal | With Our Unvoiced Excitation Signal | Change |
| Voicing | 88.0 | 83.6 | -4.4 |
| Nasality | 94.5 | 99.2 | +4.7 |
| Sustention | 74.0 | 77.1 | +3.1 |
| Sibilant | 80.2 | 84.9 | +4.7 |
| Graveness | 63.5 | 77.9 | +14.4 |
| Compactness | 88.5 | 87.8 | -0.7 |
| Overall | 81.5 | 85.1 | +3.6 |

Table 8—DRT score changes in the attribute graveness. This table lists the four features within the attribute graveness and the changes in scores when the conventional unvoiced excitation signal is replaced by our unvoiced excitation signal in the narrowband LPC.

| Features in Graveness | Feature Present | | Feature Absent | |
|-----------------------|--------------------|--------------|---------------------|--------------|
| | Test Words | Score Change | Test Words | Score Change |
| Voiced | Weed Bid Met | +14.6 | Reed Did Net | +4.2 |
| Unvoiced | Peek Fin | +17.7 | Teak Thin | +20.9 |
| Plosive | Peek Bid | +21.9 | Teak Did | +23.0 |
| Nonplosive | Weed Fin Met | +10.4 | Reed Thin Net | +2.1 |

With the conventional LPC the tendency on the DRT is for listeners to mistake unvoiced stop consonants for the voiced ones because the bursts are not reproduced well. The improved burst reproduction with the modified unvoiced excitation signal reverses this tendency—the voiced sounds are instead mistaken for unvoiced. This may be largely due to the fact that many of the plosive consonants on the original tape were articulated directly into the microphone, thus overemphasizing the bursts. Since the bursts of voiced stops are normally weaker than those of unvoiced stops, more faithful reproduction of these overly strong voiced bursts led listeners to mistakenly identify them as unvoiced. This tendency accounts for much of the drop in the "voicing" attribute score, and is consistent with the improvements produced by our modified unvoiced excitation signal.

EXPANDED OUTPUT BANDWIDTH

Since the investigation of the vocoder by Dudley in 1939, all vocoders have been implemented with the input and output bandwidths equal, and more or less confined to 4 kHz and below. This has also been true in the development of digitally implemented voice processors such as the narrowband LPC. The limited bandwidth, combined with spectral distortions caused by the low data-rate encoding, makes the synthesized speech sound rather muffled, particularly for unvoiced fricatives and stop consonants. We introduce a method of expanding the bandwidth of the synthesized speech to 6 kHz by folding the frequency contents between 2 and 4 kHz upward around the cutoff frequency of 4 kHz.

Reasons for Output Bandwidth Expansion

The primary reason for expanding the narrowband LPC output bandwidth is to allow more realistic reproduction of unvoiced speech sounds, particularly stop consonants and voiceless fricatives. We know from the spectrograms of unprocessed speech that the spectra of these sounds often extend to 6 kHz or beyond. We also know that there is little distinctive formant information in these sounds, so that the spectrum between 2 and 4 kHz is similar to that between 4 and 6 kHz. Thus, by folding the frequency contents between 2 and 4 kHz upward into the region between 4 and 6 kHz, we can make the spread of the synthesized speech similar to that of the original speech. The presence of the higher frequencies makes stop consonants sound sharper and makes voiceless fricatives sound more hissy.

The output bandwidth expansion also enhances the reproduction of voiceless fricatives whose spectra were originally above the passband of the LPC, but which were brought down within the passband by the selectively applied aliasing process described as part of our LPC analysis improvements [1]. The sound quality will be improved because the output bandwidth expansion operation is the complement of the aliasing process.

The output bandwidth expansion also allows the use of an output low-pass filter which cuts off more gently than that of the conventional narrowband LPC. If the low-pass filter cutoff is too sharp (in excess of 100 dB/octave), the unvoiced fricative tends to whistle because the cutoff frequency behaves as a resonant frequency. (Note that a sharp cutoff low-pass filter is never used in the playback of noisy 78 RPM acoustic records.) With the output bandwidth expansion, the output low-pass filter may decrease gradually from -3 dB at 4 kHz to -60 dB at 8 kHz.

The effect of the output bandwidth expansion on voiced speech is of interest, too. Unlike voiceless speech, voiced speech usually does contain formant information between 2 and 4 kHz which is reflected into the frequency range between 4 and 6 kHz by the output bandwidth expansion process. For a majority of voices, however, the intensities of the reflected formants are weak, as will be illustrated later. Even for voices with strong upper formant frequencies, the presence of the reflected formants does not affect the speech intelligibility. In fact it tends to make the synthesized speech brighter, somewhat akin to the extraneous formant frequencies of the singing voice [22], often called "singers' formants."

Finally, the expansion of the output bandwidth from 150-4000 Hz to 150-6000 Hz brings further improvement in the speech quality by shifting the tonal centroid of the LPC processed speech to a more favorable location. We feel that the conventional narrowband LPC processed speech sounds rather "bass heavy," even though the frequency components below 150 Hz are attenuated at the rate of 18 dB/octave. We do not get a similar feeling with unprocessed speech, even though none of the low-frequency components are attenuated. One explanation for this effect is that the tonal centroid of the unprocessed speech is located at a higher frequency because the bandwidth extends to 10 kHz or above. Similarly the expansion of the output bandwidth in the LPC helps raise the tonal centroid of the processed speech. Note that raising the lower cutoff frequency produces a similar perceptual effect; for this reason the lower cutoff frequency for the telephone is often as high as 300 Hz. However, the lower cutoff frequency of the narrowband LPC cannot be much higher than the current 150 Hz because both the pitch tracking and the voicing decision would suffer greatly when another narrowband LPC is operated in tandem, as often happens in military communication setups.

Output Bandwidth Expansion Process

The output bandwidth expansion process is a two-step postsynthesis operation on the speech samples synthesized by the narrowband LPC. The two steps required are (a) the spectral folding process at double the sampling rate and (b) a low-pass filtering operation. Each is discussed below.

Spectral Folding Process

The spectral folding process simply involves adding a zero between every pair of adjacent samples in the output digital waveform. The input and output of the spectral folding process are depicted in Figs. 17(a) and 17(b) respectively. As noted, the sampling frequency is increased by a factor of 2.

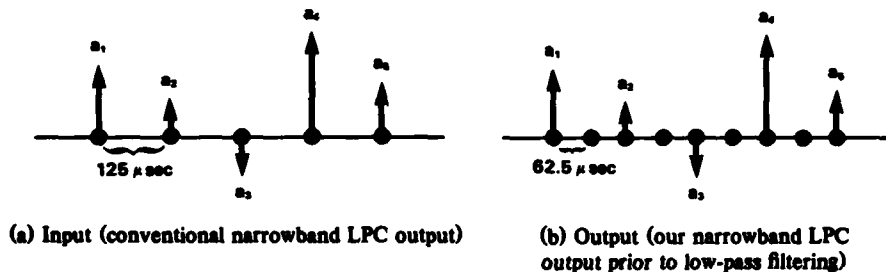


Fig. 17 — Input and output of the spectral folding process in the sampled-data form.
Note that the sampling rate is doubled at the output.

The sampled data representation of the conventional LPC output is expressed in the form

$$X(z) = \sum_{n=-\infty}^{\infty} x(nT)z^{-n} \quad (21)$$

where $x(nT)$ is the n th sampled value of the synthesized speech, T is the sampling time interval ($125 \mu\text{s}$ for the conventional narrowband LPC using an 8 kHz sampling rate), and z^{-1} is the shifting operation by one sampling time interval. The spectrum of the sampled signal is the sum of the signal spectrum from 1 to $1/2T$ Hz and the shifted complementary spectra at every multiple of $1/T$ Hz.

The sampled data representation of the output of the spectral folding process may be expressed by

$$X(z) = \begin{cases} \sum_{n=-\infty}^{\infty} x\left(\frac{nT}{2}\right) z^{-n/2} & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases} \quad (22)$$

or

$$X(z) = \frac{1}{2} \left[\sum_{n=-\infty}^{\infty} x\left(\frac{nT}{2}\right) z^{-n/2} + \sum_{n=-\infty}^{\infty} x\left(\frac{nT}{2}\right) (-1)^n z^{-n/2} \right]. \quad (23)$$

Within the expanded passband of 0 to $1/T$ (i.e., 0 to 8 kHz), the signal spectrum from 0 to $1/2T$ (0 to 4 kHz) is contributed by the first term in Eq. (23), and the folded spectrum from $1/2T$ to $1/T$ (4 to 8 kHz) is contributed by the second term.

Low-Pass Filtering

This may be accomplished by a single analog filter at the output of the digital-to-analog (D-A) converter, or by a combination of a digital filter prior to the D-A converter and a less stringently designed analog filter at the output of the D-A converter. We used the first approach in a real-time implementation by using the existing LPC processor, and the second approach in a nonreal-time simulation. The filter characteristics are not too critical, but we recommend that the attenuations at 4 and 8 kHz should be about 3 and 60 dB, respectively.

Spectrographic Analyses of Narrowband LPC Output

This output bandwidth expansion process has been incorporated in our conventional narrowband LPC operating in real time. We used an analog filter to suppress the frequency contents above 6 kHz, since the computational time available from the processor was insufficient for digital filtering. Figure 18 shows spectrographic analyses of a female voice before and after LPC analysis and synthesis. As noted, the fricative sounds at the end of THOSE and the beginning of CHILDREN are reproduced beyond the 4 kHz passband of the LPC. The resulting sound quality is noticeably closer to that of the unprocessed speech. Even the small burst waveform at the onset of DIRTY has been reproduced with an expanded bandwidth as in the unprocessed speech. Figure 19 presents similar spectrographic analyses of a male voice.

Figure 20 is the narrowband spectrographic analysis of the same female voice shown in Fig. 18. Note that the pitch harmonics are evenly spaced in the frequency range above 4 kHz, indicating that no audible distortions are created by noninteger pitch harmonics.

Test and Evaluation

The use of the extended output bandwidth makes the synthesized LPC speech noticeably brighter, less muffled, and more pleasant to listen to. To evaluate this improvement quantitatively, we turned once again to the DAM test. The results show a 2.5-point increase in the overall quality score from 46.7 for the conventional LPC to 49.2 with the extended output bandwidth (3 male and 3 female speakers).

A DRT was also run to evaluate the effect of the extended output bandwidth on the intelligibility of the LPC speech. As we expected, the DRT score for our modified LPC was virtually identical to that of the conventional LPC, indicating that this modification does improve the speech quality with no adverse effect on the speech intelligibility.

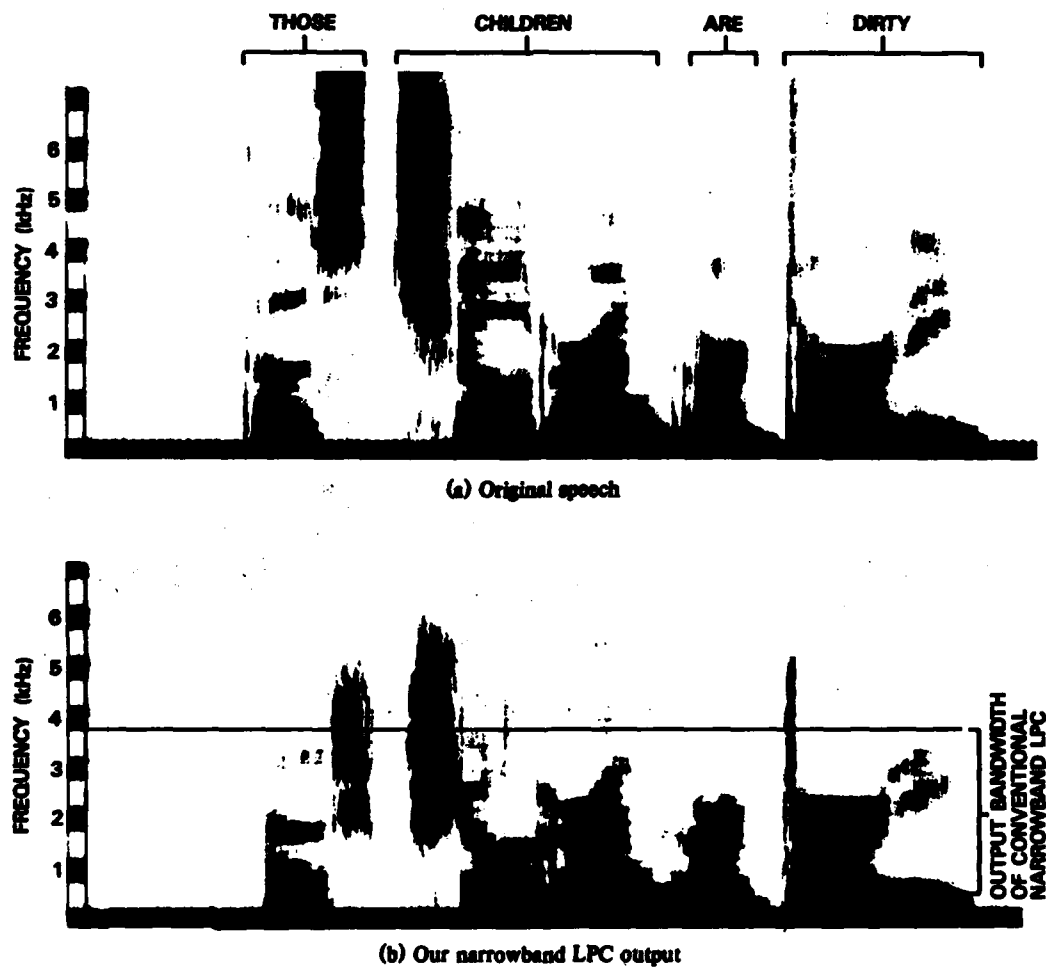
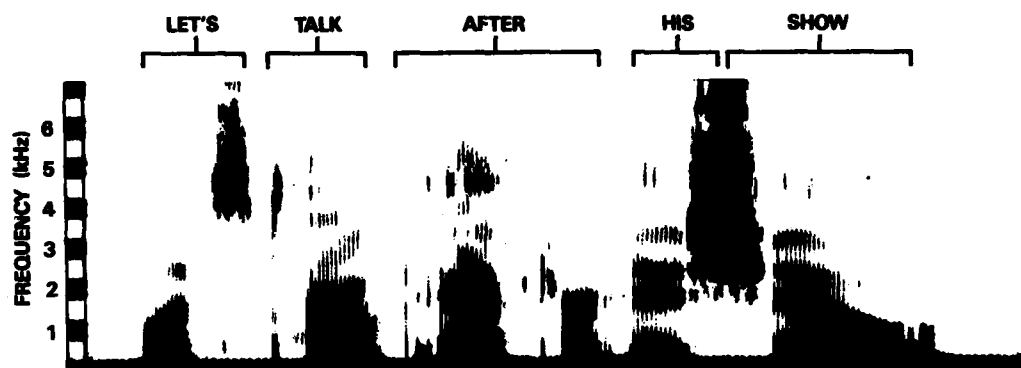
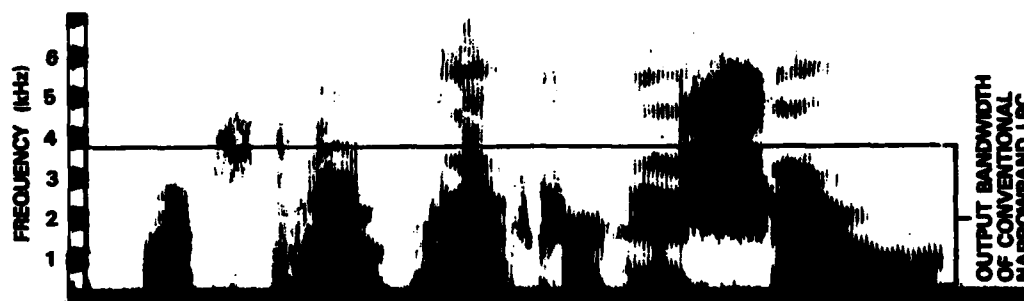


Fig. 18 — Spectrographic analysis of original female speech and the output of the narrowband LPC with our bandwidth expansion. Note that the fricative spectra are spread beyond the passband of the conventional narrowband LPC. Since their spectral distributions are more similar to those of the original speech they sound more natural.



(a) Original speech



(b) Our narrowband LPC output

Fig. 19 — Spectrographic analysis of original male speech and output of the narrowband LPC with our bandwidth expansion

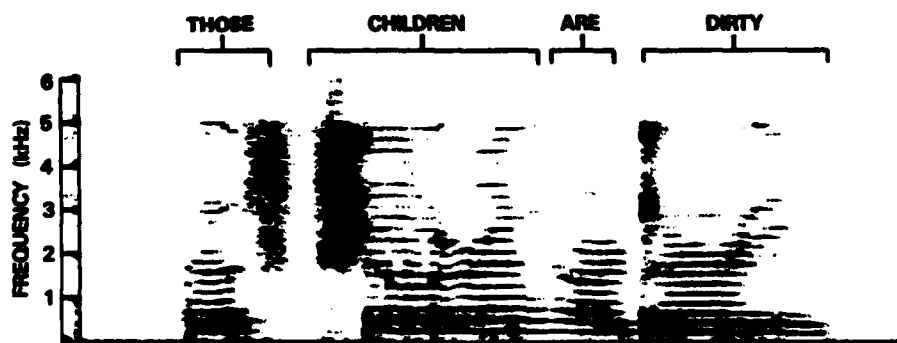


Fig. 20 — Narrowband spectrographic analysis of the LPC output with our bandwidth expansion (female speech). Evenly spaced pitch harmonics indicate that the output bandwidth expansion process does not introduce pitch deformations. The wideband spectrographic analysis of the same speech is shown in Fig. 18(b).

CONCLUSIONS

The objective of this effort was to improve the narrowband LPC speech without compromising the existing DoD interoperability requirements on the speech sampling rate, the frame rate, and the parameter coding formats. These requirements are expected to remain unaltered for many years. Thus, it is essential to work within these constraints so that any useful results from these research efforts will benefit the Navy and DoD in general.

Since the narrowband LPC transmits the speech at a low bit rate (less than 5% of the original speech transmission rate), some of the speech parameters—particularly those of the excitation source—cannot be transmitted and are introduced at the receiver as fixed parameters. The major weakness of the narrowband LPC synthesizer lies in the use of fixed parameters which do not reflect the changing nature of human speech. We have modified the amplitude and phase spectra of the voiced excitation signal, as well as the temporal characteristics of the unvoiced excitation, to simulate some of these natural irregularities.

Though these modifications can be implemented independently, the greatest benefit is obtained when these synthesis improvements are combined with the analysis improvements we presented in an earlier report. The speech quality and intelligibility of the resulting narrowband LPC will be nearly comparable to that of a voice processor operating at four times the data rate of the LPC.

After nearly a decade of research and development, the narrowband LPC is now a practical means for digitizing speech at low bit rates and is becoming widely deployed in military platforms and communication centers. Our efforts, and similar efforts by other investigators, will help make the narrowband LPC more acceptable to general users.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the NRL Research Advisory Committee. We thank Robert Martin and Jack Garner of the Naval Electronic Systems Command, who supported other activities related to this effort; in particular they supported Mark Lidd of Signal Processing Solutions, Inc., who generated the software for the real-time implementations to test some of our ideas. We also thank Professor Matthew Yuschik of the University of South Carolina, who spent the summer of 1982 with us. Our numerous discussions on the speech synthesis process were most enlightening.

Finally, we appreciate the help of Ruth Phillips of NRL, who maintained our computer facilities in good working order. Our effort could not have been completed without her help.

REFERENCES

1. G.S. Kang and S.S. Everett, "Improvement of the Narrowband Linear Predictive Coder, Part 1—Analysis Improvements," NRL Report 8645, Dec. 1982.
2. W.D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," in *Conf. Rec., 1977 IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, May 1977, pp. 204-207.
3. W.D. Voiers, "Diagnostic Evaluation of Speech Intelligibility," in *Speech Intelligibility and Recognition*, M.E. Hawley, ed. (Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1977).
4. G.S. Kang and L.J. Fransen, "Second Report of the Multirate Processor (MRP) for Digital Voice Communications," NRL Report 8614, Sept. 1982.

5. B.S. Atal and J.R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," in *Conf. Rec., 1982 IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, May 1982, pp. 614-617.
6. T.E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology* 1 (2), 40-49 (Apr. 1982).
7. G.S. Kang, L.J. Fransen and E.L. Kline, "Multirate Processor (MRP) for Digital Voice Communication," NRL Report 8295, Mar. 1979, Appendix C.
8. G.S. Kang, "Application of Linear Predictive Encoding to a Narrowband Voice Digitizer," NRL Report 7774, Oct. 1974.
9. M.R. Schroeder, "Synthesis of Low-Peak-Factor Signals and Binary Sequences with Low Auto-correlation," *IEEE Trans. Inf. Theory*, IT-16, 85-89 (Jan. 1970).
10. M.R. Sambur, A.E. Rosenberg, L.R. Rabiner, and C.A. McGonegal, "On Reducing the Buzz in LPC Synthesis," *Jour. Acoust. Soc. Am.* 63, 918-924, Mar. 1978.
11. F.L. Wightman and D.M. Green, "The Perception of Pitch," *Sci. Am.* 22, 208-215 (1974).
12. B.S. Atal and N. David, "On Finding the Optimum Excitation for LPC Speech Synthesis," *Jour. Acoust. Soc. Am.* 63, Supp. 1, S79 (Spring 1978).
13. B.S. Atal and N. David, "On Synthesizing Natural-Sounding Speech by Linear Prediction," in *Conf. Rec., 1979 IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Apr. 1979, pp. 44-47.
14. J. Makhoul, R. Viswanathan, R. Schwartz, and A.W.F. Huggins, "A Mixed-Source Model for Speech Compression and Synthesis," *Jour. Acoust. Soc. Am.* 64, 1577-1581 (Dec. 1978).
15. O. Fujimura, "An Approximation to Voice Aperiodicity," *IEEE Trans. Audio Electroacoust.* AU-16, 68-72 (Mar. 1968).
16. D. Coulter, "Application of Simultaneous Voice/Unvoice Excitation in a Channel Vocoder," U.S. Patent 3,903,366, 1975.
17. J.N. Holmes, "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer," *IEEE Trans. Audio Electroacoust.* AU-21, 298-305 (June 1973).
18. L.R. Rabiner and R.E. Crochiere, "On the Design of All-Pass Signals with Peak Amplitude Constraints," *Bell Sys. Tech Jour.* 55, 395-407 (Apr. 1976).
19. D.C. Coulter, C.L. Ludlow, and F.H. Gentges, "Limits of Frequency Perturbation in Normal Phonation," *Jour. Acoust. Soc. Am.*, in press.
20. P. Stevens, "Spectra of Fricative Noise in Human Speech," *Language and Speech* 3, 202-219 (1960).
21. F. Minifie, T. Hixon and F. Williams, *Normal Aspects of Speech, Hearing and Language* (Prentice Hall, Englewood Cliffs, N.J., 1973).
22. J. Sunberg, "The Acoustics of the Singing Voice," *Scientific American*, 82-91 (Mar. 1977).

END

FILMED

9-85

DTIC